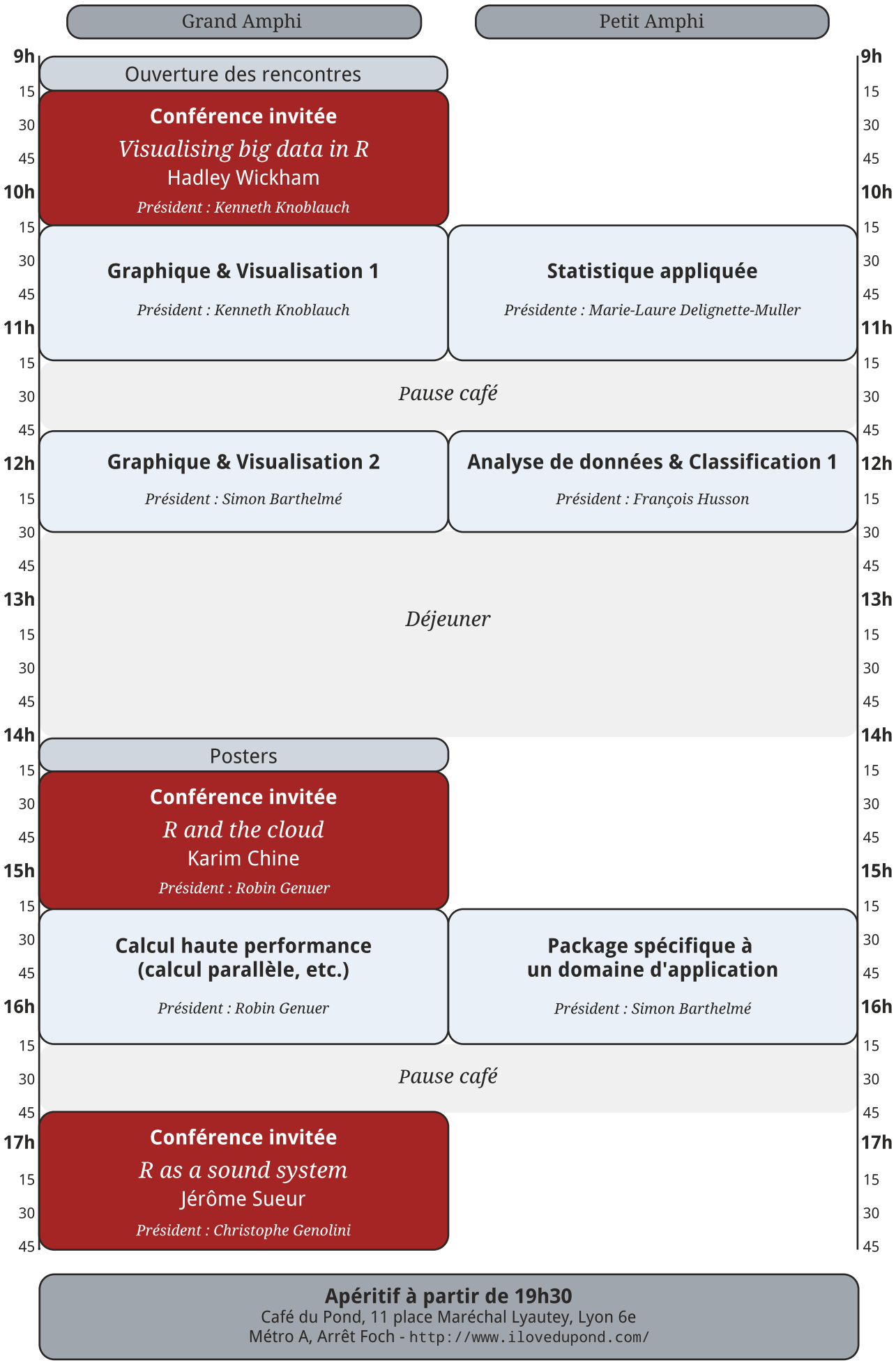
Deuxièmes rencontres **R**

Lyon, 27 et 28 juin 2013

Programme
Recueil des résumés

http://r2013-lyon.sciencesconf.org/

Jeudi 27 juin 2013 - Programme



Jeudi 27 juin 2013 - Descriptif des sessions

Grand Amphi / 10:15-11:15

Graphique & Visualisation 1

10:15-10:35
François Husson
L'analyse de données avec FactoMineR : les nouveautés

10:35-10:55
Timothée Giraud
rCarto, un package de cartographie statistique

10:55-11:15
Aurélie Siberchicot
adegraphics : un package pour la représentation et l'analyse de données multivariées

Petit Amphi / 10:15-11:15

Statistique appliquée

10:15-10:35
Delphine Rieutort
Développement d'une application sous R pour la Surveillance Observationnelle des Problèmes de Santé au Travail

10:35-10:55
Julien Clavel
Estimation des données manquantes en morphométrie : quelle limite choisir ?

10:55-11:15
Jeremie Zaffran
Prédiction de la réactivité du glycérol sur les catalyseurs métalliques

Grand Amphi / 11:45-12:25

Graphique & Visualisation 2

11:45-12:05
Kenneth Knoblauch
What a statistician might want to know about human color vision, but was afraid to ask !

12:05-12:25
Christophe Genolini
KmL3D : K-means pour données longitudinales jointes

Petit Amphi / 11:45-12:25

Analyse de données & Classification 1

11:45-12:05
Amaury Labenne
MFAMIX : une extension de l'analyse factorielle multiple pour des groupes de variables mixtes

12:05-12:25
Jean Thioulouse
Méthodes de couplage de deux K-tableaux et collections de graphiques

Grand Amphi / 15:15-16:15

Calcul haute performance (calcul parallèle, etc.)

15:15-15:35
Habib Saadi
R2GUESS: a GPU-based R package for sparse Bayesian variable selection

15:35-15:55
Didier Chauveau
An R package using HPC for entropy estimation and MCMC evaluation

15:55-16:15
Romain Francois
Intégration R et C++ avec Rcpp

Petit Amphi / 15:15-16:15

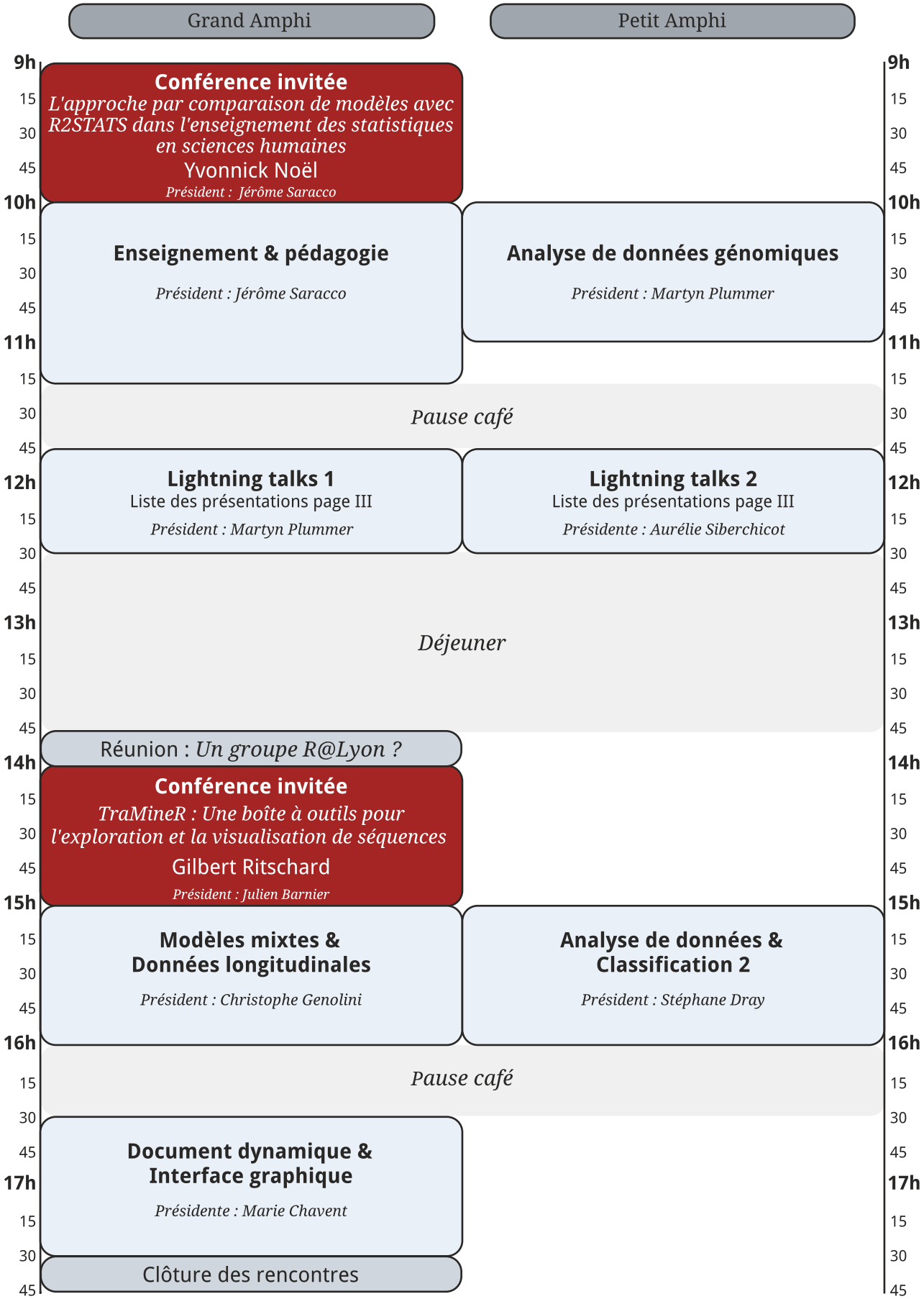
Package spécifique à un domaine d'application

15:15-15:35
Alexandre Laurent
frailtypack : Un package pour l'analyse de données de survie corrélées

15:35-15:55
Emmanuel Rousseaux
The Dataset Project: Handling survey data in R

15:55-16:15
Jeremie Riou
Sample Size Determination and Data Analysis in the context of continuous co-primary endpoints in clinical trials

Vendredi 28 juin 2013 - Programme



Vendredi 28 juin 2013 - Descriptif des sessions

Grand Amphi / 10:00-11:20

Enseignement & pédagogie

10:00-10:20 / Simon Penel
Un site web d'enseignement R automatisé et à gestion partagée

10:20-10:40 / Sylvain Mousset
Génération automatique de documents pédagogiques avec R pour l'enseignement et l'évaluation des étudiants

10:40-11:00 / Mehdi Khaneboubi
Pistes de réflexion pour la mise en œuvre d'un enseignement à distance sur le test du khi carré d'indépendance pour des étudiants en master de sciences de l'éducation

11:00-11:20 / Anne-Béatrice Dufour
De la biologie à l'algèbre linéaire ... en passant par R. Expérimenter la notion de projection

Petit Amphi / 10:00-11:00

Analyse de données génomiques

10:00-10:20 / Guillemette Marot
metaRNASeq: un package pour la méta-analyse de données RNASeq

10:20-10:40 / Morgane Pierre-Jean
jointSeg : Segmentation de données génomiques en cancérologie

10:40-11:00 / Andrea Rau
HTSFilter: An independent data-based filter for replicated high-throughput transcriptome sequencing experiments

Grand Amphi / 15:00-16:00

Modèles mixtes & Données longitudinales

15:00-15:20 / Viviane Philipps
multlcmm : fonction d'estimation de modèles mixtes à processus latent pour données longitudinales

15:20-15:40 / Cécile Sauder
Prédiction d'un événement binaire à partir de données fonctionnelles : application aux bovins laitiers

15:40-16:00 / Djeneba Thiam
Données tronquées sous R dans les modèles linéaires simples et à effets mixtes

Petit Amphi / 15:00-16:00

Analyse de donnée & Classification 2

15:00-15:20 / Quentin Grimonprez, Julien Jacques
Rankclust: An R package for clustering multivariate partial rankings

15:20-15:40 / Nathalie Villa-Vialaneix
SOMbrero: Cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités

15:40-16:00 / Matthieu Cornec
The data.sample package: sampling as a big data mining tool

Grand Amphi / 16:30-17:30

Document dynamique & Interface graphique

16:30-16:50 / David Gohel
Génération de documents Word à partir de R : utilisation du package R2DOCX dans une plateforme statistique en milieu industriel

16:50-17:10 / Damien Leroux
sexy-rgtk: a package for programming RGtk2 GUI in a user-friendly manner

17:10-17:30 / Rémy Drouilhet
R-dyndoc une alternative à Sweave

Table des matières

Jeudi 27 juin 2013 - 09:15 - 10:15

Grand Amphi : Conférence invitée

Visualising big data in R, H. Wickham 1

Jeudi 27 juin 2013 - 10:15 - 11:15

Grand Amphi : Graphique & Visualisation 1

L'analyse de données avec FactoMineR : les nouveautés, F. Husson 2

rCarto, un package de cartographie statistique, T. Giraud 4

adegraphics : un package pour la représentation et l'analyse de données multivariées, A. Siberchicot [et al.] 6

Petit Amphi : Statistique appliquée

Développement d'une application sous R pour la Surveillance Observationnelle des Problèmes de Santé au Travail, D. Rieutort [et al.] 8

Estimation des données manquantes en morphométrie : quelle limite choisir ?, J. Clavel [et al.] 10

Prédiction de la réactivité du glycérol sur les catalyseurs métalliques, J. Zaffran [et al.] 12

Jeudi 27 juin 2013 - 11:45 - 12:25

Grand Amphi : Graphique & Visualisation 2

What a statistician might want to know about human color vision, but was afraid to ask!, K. Knoblauch 14

KmL3D: K-means pour données longitudinales jointes, C. Genolini [et al.] 16

Jeudi 27 juin 2013 - 11:45 - 12:45

Petit Amphi : Analyse de données & Classification 1

MFAMIX : une extension de l'analyse factorielle multiple pour des groupes de variables mixtes, A. Labenne 18

Méthodes de couplage de deux K-tableaux et collections de graphiques, J. Thioulouse [et al.] 20

ACP Fonctionnelles de Densités de Probabilité Estimées par la Méthode du Noyau Multivariée avec R, S. Yousfi [et al.] 22

Jeudi 27 juin 2013 - 14:00 - 14:15

Grand Amphi : Poster

aste: An R package for the adaptive estimation the right tail, F. Caeiro 24

Cascade : un package R pour étudier la dispersion d'un signal dans un réseau de gènes., N. Jung [et al.] 26

Corrélations entre maillages 3D au moyen du logiciel R application à l'imagerie par IRM de l'accident vasculaire cérébral, A. Rouanet [et al.] 29

Corrélations entre signaux EEG : un code d'analyse parallélisé, A. Cheylus [et al.] 31

Parallel Computing in R using the Bot Package, F. Chuffart 32

Un package pour utiliser les Cumulative Distribution Networks, V. Pham [et al.] 33

Visualisation de processus spatiaux à l'aide de la correction de Ripley, A. Charpentier [et al.] 34

Visualisation et cartographie des données de capteurs météorologiques à l'échelle des terroirs viticoles, M. Madelin [et al.] 35

Jeudi 27 juin 2013 - 14:15 - 15:15

Grand Amphi : Conférence invitée

R and the Cloud, K. Chine 36

Jeudi 27 juin 2013 - 15:15 - 16:15

Grand Amphi : Calcul haute performance (calcul parallèle, etc.)

R2GUESS: a GPU-based R package for sparse Bayesian variable selection, H. Saadi 38

An R package using HPC for entropy estimation and MCMC evaluation, D. Chauveau [et al.] 40

Intégration R et C++ avec Rcpp, R. Francois 42

Petit Amphi : Package spécifique à un domaine d'application

frailtypack : Un package pour l'analyse de données de survie corrélées, A. Laurent [et al.] 43

The Dataset Project: Handling survey data in R, E. Rousseaux [et al.] 45

Sample Size Determination and Data Analysis in the context of continuous co-primary endpoints in clinical trials., J. Riou 47

Jeudi 27 juin 2013 - 16:45 - 17:45

Grand Amphi : Conférence invitée

R as a sound system, J. Sueur 49

Vendredi 28 juin 2013 - 09:00 - 10:00

Grand Amphi : Conférence invitée

L'approche par comparaison de modèles avec R2STATS dans l'enseignement des statistiques en sciences humaines, Y. Noel 51

Vendredi 28 juin 2013 - 10:00 - 11:20

Grand Amphi : Enseignement & Pédagogie

Un site web d'enseignement R automatisé et à gestion partagée, S. Penel [et al.] 53

Génération automatique de documents pédagogiques avec R pour l'enseignement et l'évaluation des étudiants., S. Mousset 55

Pistes de réflexion pour la mise en oeuvre d'un enseignement à distance sur le test du khi carré d'indépendance pour des étudiants en master de sciences de l'éducation, M. Khaneboubi 56

De la biologie à l'algèbre linéaire ... en passant par R. Expérimenter la notion de projection., A. Dufour [et al.] 58

Vendredi 28 juin 2013 - 10:00 - 11:00

Petit Amphi : Analyse de données génomiques

metaRNASeq: un package pour la méta-analyse de données RNASeq, G. Marot [et al.] 60

jointSeg : Segmentation de données génomiques en cancérologie, M. Pierre-jean 62

HTSFilter: An independent data-based filter for replicated high-throughput transcriptome sequencing experiments, A. Rau [et al.] 64

Vendredi 28 juin 2013 - 11:45 - 12:30

Grand Amphi : Lightning talks 1

Aspects de cartographie thématique pour les sciences sociales avec R, J. Gombin ... 66
Nouvelles fonctionnalités du package fitdistrplus, M. Delignette-muller [et al.] ... 67
SesIndexCreatoR : Un package R pour la création et la visualisation d'indices socioéconomiques, B. Lalloué [et al.] ... 69
PNN : une nouvelle bibliothèque R pour la modélisation d'un réseau de neurones probabilistes de Specht, P. Chasset ... 71
Estimation des risques dans les formes familiales de cancer., Y. Drouet [et al.] ... 73
Prévision de consommation électrique avec R, R. Nedellec ... 75
autoplot : ready-made plots with ggplot2, J. Irisson ... 76

Petit Amphi : Lightning talks 2

Des analyses exploratoires multidimensionnelles pour prédire la progression des patients en thérapie, T. Delespierre [et al.] ... 78
R au secours des écotoxicologues, G. Kon kam king [et al.] ... 80
Analyse et datation d'artefacts archéolgiques: R et les cachets circulaires hittites, N. Strupler ... 82
Teaching R to social science undergraduates, F. Briatte [et al.] ... 84
R tools for spatial point pattern analysis applied to fluorescence localization nanoscopy, J. Godet [et al.] ... 85
Traiter des données de tracking ? navigation web, achats, suivi d'enquête ? avec R, A. Gayet ... 86
Rendre R plus accessible aux utilisateurs non-informaticiens, J. Maalouf [et al.] ... 88

Vendredi 28 juin 2013 - 14:00 - 15:00

Grand Amphi : Conférence invitée

TraMineR : Une boite à outils pour l'exploration et la visualisation de séquences, G. Ritschard ... 89

Vendredi 28 juin 2013 - 15:00 - 16:00

Grand Amphi : Modèles mixtes & Données longitudinales

multlcmm : fonction d'estimation de modèles mixtes à processus latent pour données longitudinales, V. Philipps [et al.] ... 91
Prédiction d'un événement binaire à partir de données fonctionnelles : Application aux bovins laitiers, C. Sauder [et al.] ... 93
Données tronquées sous R dans les modèles linéaires simples et à effets mixtes, D. Thiam [et al.] ... 95

Petit Amphi : Analyse de données & Classification 2

Rankclust: An R package for clustering multivariate partial rankings, Q. Grimonprez [et al.] ... 97
SOMbrero: Cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités, L. Bendhaïba [et al.] ... 99
The data.sample package: sampling as a big data mining tool, M. Cornec [et al.] ... 101

Vendredi 28 juin 2013 - 16:30 - 17:30

Grand Amphi : Document dynamique & Interface graphique

Génération de documents Word à partir de R : utilisation du package R2DOCX dans une plateforme statistique en milieu industriel., D. Gohel [et al.] ... 102
sexy-rgtk: a package for programming RGtk2 GUI in a user-friendly manner, D. Leroux [et al.] ... 103

R-dyndoc une alternative à Sweave, R. Drouilhet 105

Visualising big data in R

H. Wickham[a]

RStudio
h.wickham@gmail.com

Mots clefs : Visualisation, big data

R has a notorious reputation for not being able to deal with "big" data (and ggplot2 is a frequent culprits). Fortunately, this isn't an underlying problem with R, and it's something that we can fix with good programming practices and intelligent use of compiled code. In this talk, I'll introduce a new package, bigvis, that aims to make it easier (and faster) to work with very large datasets.
Bigvis makes it possible to visualise 10-100 million observations in just a few seconds. It is built around a pipeline of bin, summarise, smooth and visualise, and makes minimal sacrifices of flexibility to achieve fast performance. As well as discussing the visualisation challenges when you have 10s of millions of observations, I'll also discuss the performance challenges, and how C++ and Rcpp make it pleasurable to integrate compiled code into R.

L'analyse de données avec FactoMineR : les nouveautés

François Husson et **Julie Josse**

Laboratoire de Mathématiques appliquées
Agrocampus Ouest
65 rue de Saint-Brieuc - 35042 Rennes
husson@agrocampus-ouest.fr
josse@agrocampus-ouest.fr

Mots clefs : Analyse de données, module graphique, `FactoMineR`, données manquantes.

Plusieurs packages sont disponibles pour faire de l'analyse des données sous `R`, citons par exemple les packages `ade4`, `FactoMineR` et `ca`. Nous nous focalisons dans cet exposé sur le package `FactoMineR` [1] et nous présentons les dernières nouveautés, notamment le module graphique et la façon dont les données manquantes sont gérées.
Les tableaux de données incomplets sont parfois gérés de façon très succincte dans les logiciels : la donnée manquante est remplacée par la moyenne pour les données quantitatives ou encore la donnée manquante génère une modalité que l'on peut appeler "manquante" quand les données sont qualitatives. Ceci est une première approche d'imputation des données manquantes mais [2] ont proposé de nouvelles méthodes d'imputation qui sont disponibles dans le package `missMDA` [3]. Ce package peut être utilisé en complément du package `FactoMineR` pour faire des analyses en composantes principales (ACP), des analyses des correspondances multiples (ACM), des analyses factorielles de données mixtes (AFDM) ou des analyses factorielles multiples (AFM) avec données manquantes.
En ce qui concerne le module graphique de FactoMineR, il possède maintenant un algorithme qui positionne de façon "optimale" les libellés des éléments (individus, variables, modalités, fréquences, groupes de variables ... selon la méthode utilisée, ACP, analyse des correspondances (AFC), ACM, AFDM, AFM). L'algorithme calcule et minimise un taux de recouvrement entre libellés ce qui permet d'avoir des graphes plus lisibles. Il est également possible de colorier les libellés en fonction d'une variable qualitative, de ne mettre les libellés que de certains points : par exemple ceux qui ont le plus contribué à la construction du plan factoriel, ceux qui sont projetés avec une qualité de représentation suffisante, ou encore de les sélectionner par leur nom. Là encore, cela permet d'améliorer la lisibilité des graphes. La figure 1' montre un graphe d'ACP obtenu en coloriant les individus en fonction d'une variable qualitative à 2 modalités (athlètes participant à un décathlon lors d'un Decastar ou lors de Jeux Olympiques). Seuls les individus ayant une bonne qualité de représentation dans le plan (cosinus carré supérieur à 0.6) sont coloriés, les autres individus sont positionnés mais dessinés avec une certaine transparence et sans libellé). Le graphe 1 correspond au graphe directement obtenu avec les lignes de code suivantes :

```
> library(FactoMineR)
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13)
> plot(res.pca,habillage="Competition",select="cos2 0.6")
```

Cette sélection des libellés les plus intéressants qui permet de conserver la vision de l'ensemble du nuage de points est très utile pour les graphes ayant beaucoup d'éléments, comme par

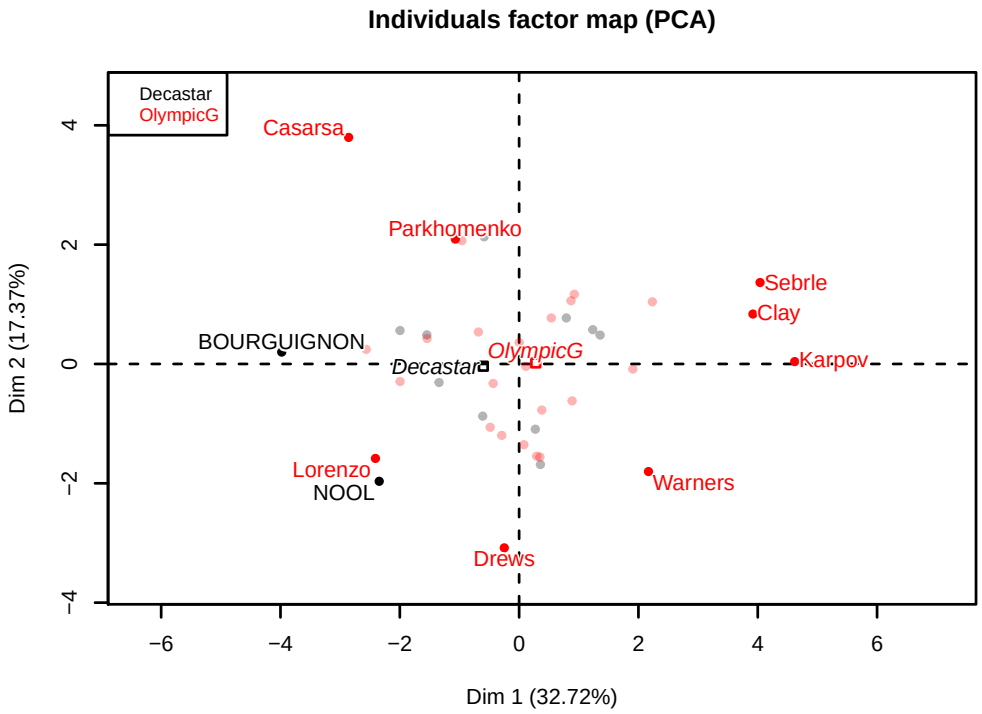


FIGURE 1 – Graphe des individus obtenu directement par `FactoMineR`. Les individus sont coloriés en fonction d'une variable qualitative à 2 modalités et seuls les individus ayant une qualité de projection suffisante (cosinus carré supérieur à 0.6) ont un libellé.

exemple pour l'analyse d'une enquête par ACM ou par l'analyse de données textuelles par AFC. Des vidéos disponibles sur Youtube en français ou en anglais permettent de voir comment utiliser le module graphique, comment gérer les données manquantes, etc.

Références

[1] Husson, F., Josse, J., Lê, S. & Mazet, J. (2013). FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R, R package version 1.24, `http://factominer.free.fr`.
[2] Josse, J. & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), p 79-99.
[3] Husson, F. & Josse, J. (2013). missMDA : Handling missing values with/in multivariate data analysis (principal component methods), R package version 1.7.

rCarto, un package de cartographie statistique

Timothée Giraud [a]

[a] UMS RIATE
Université Paris Diderot – CNRS - Datar
Université Paris 7 – UFR GHSS Case 7001 – 75025 Paris cedex 13
timothee.giraud@ums-riate.fr

Mots clefs : Géographie, Cartographie, Sémiologie graphique, Visualisation, Carte.

Une rapide recherche du terme « *maps* » dans l'agrégateur de blogs « R-bloggers » renvoie à de nombreux billets comprenant des cartes produites avec R, ou même des tutoriaux complets. L'intérêt pour la cartographie semble assez fort dans la communauté des utilisateurs. L'utilisation de R permet la conception automatisée de cartes tout en minimisant les ruptures dans la chaîne logicielle partant des données et aboutissant à la carte.
Après une rapide revue critiques des solutions régulièrement employées, nous présenterons un exemple de cartographie statistique respectueuse de la sémiologie graphique et comprenant les éléments d'habillage indispensables à ce type de représentation [1]. Nous aborderons plus précisément le package rCarto qui propose plusieurs types de cartes statistiques et des exemples permettrons d'explorer dans le détail les paramètres des fonctions disponibles dans ce package.

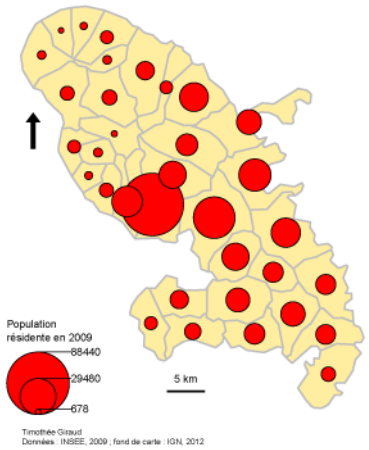
Les cartes conçues avec R sont plus ou moins sophistiquées et de natures variées. Quelques exemples de réalisation font apparaitre des points forts et des points faibles de scripts et packages actuellement utilisés, qu'ils concernent la carte elle-même ou son habillage (dimension primordiale à la bonne compréhension d'une carte).

De précédents travaux [2] [3] explorent les principes sous-jacents de la fabrication de cartes dans R. Ils s'appuient pour la gestion des données spatiales sur le package maptools permettant la manipulation des formats spécifiques aux données spatiales. Pour les traitements statistiques en amont de la représentation (l'étape de la discrétisation, primordiale dans le processus de création de la carte), nous utilisons le package classInt. Sachant que la qualité et l'efficacité visuelle d'une carte dépendent en grande partie du choix des couleurs utilisées, nous utilisons donc le package RColorBrewer issu de recherches sur l'efficacité des gammes de couleurs en cartographie.

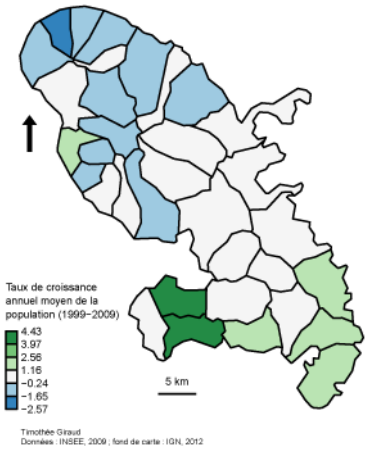
Ces premières explorations ont été rassemblées, étoffées et solidifiées dans le package rCarto dont l'objectif est de produire des cartes directement publiables, ne nécessitant pas de retouches dans un logiciel d'édition vectoriel.
Toutes les cartes proposées dans ce package sont complètes dans le sens ou elles associent à la carte elle-même l'ensemble des attributs de l'habillage d'une carte thématique complète : titre, légende détaillée (plusieurs placements possibles), échelle et orientation (placement interactif), source et auteur.
Les cinq types de cartes disponibles pour l'instant sont : les cartes de cercles proportionnels à une variable de stock (carte 1), les cartes choroplèthes (une carte choroplèthe est une carte thématique composée par la juxtaposition d'aplats de couleurs) colorées selon la discrétisation d'une variable quantitative (carte 2), les cartes de cercles proportionnels à une variable de stock et colorés selon la discrétisation d'une variable quantitative (carte 3), les cartes choroplèthes colorées selon les modalités d'une variable qualitative (carte 4), les cartes de

cercles proportionnels à une variable de stock et colorés selon les modalité d'une variable qualitative (carte 5).
Chacune des fonctions du package permet la modification de plusieurs paramètres (dimension de la figure, taille des textes, position de la légende, gammes de couleurs utilisée).

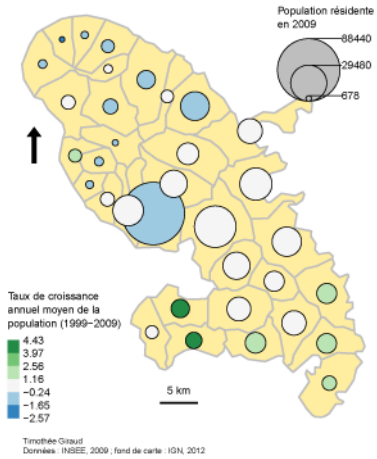
Carte 1
Distribution de la population en Martinique



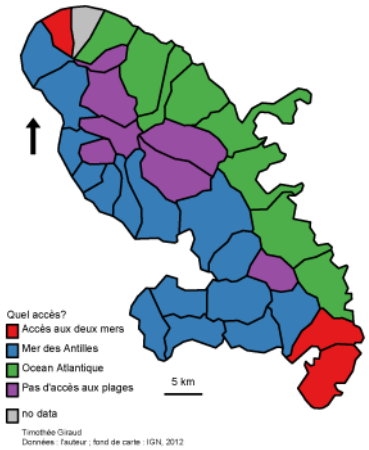
Carte 2
Dynamique de la population en Martinique



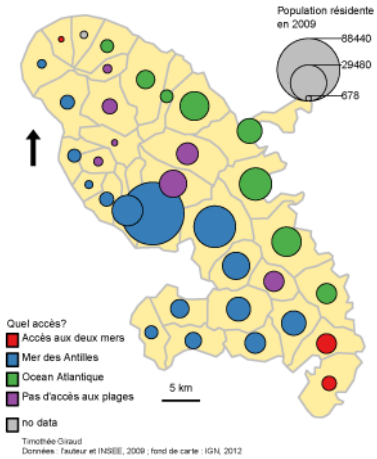
Carte 3
Dynamique de la population en Martinique



Carte 4
Accès aux plages martiniquaises



Carte 5
Accès aux plages martiniquaises



Ce package est déposé sur le CRAN. Il est en cours de développement (version 0.8) et proposera prochainement d'autres types de représentations telles que la superposition de cartes choroplèthes et de cercles proportionnels (colorés ou non).
Une page web associée à cette communication [4] propose une rapide revue commentée de billets de blogs concernant la cartographie. Les exemples proposé sont diffusés avec leurs scripts commentés et les données associées (page générée avec le package knitr).

Références

[1] Beguin, M., Pumain, D. (2000), La représentation des données géographiques : statistiques et cartographie, coll. Cursus Armand Colin, 2e édition
[2] Beauguitte, L., Giraud, T. (2012). Cartographier avec le logiciel R, http://quanti.hypotheses.org/795/
[3] Giraud, T., Severo, M. (2012). Visualisation de données médiatiques et géographiques avec R. Représenter et visualiser des données avec R - Séminaire Ined SMS, http://wukan.ums-riate.fr/
[4] Giraud, T. (2013). rCarto, un package de cartographie statistique, Rencontres R, Lyon, http://wukan.ums-riate.fr/rencontres-r/

adegraphics : un package pour la représentation et l'analyse de données multivariées

A. Siberchicot[a], **A. Julien-Laferrière**[a], **J. Thioulouse**[a] and **S. Dray**[a]

[a]Laboratoire de biométrie et biologie évolutive (UMR CNRS 5558)
CNRS - Université Lyon 1
43 bd du 11 novembre 1918, 69622 Villeurbanne, France
{aurelie.siberchicot, alice.julien-laferriere, jean.thioulouse, stephane.dray}@univ-lyon1.fr

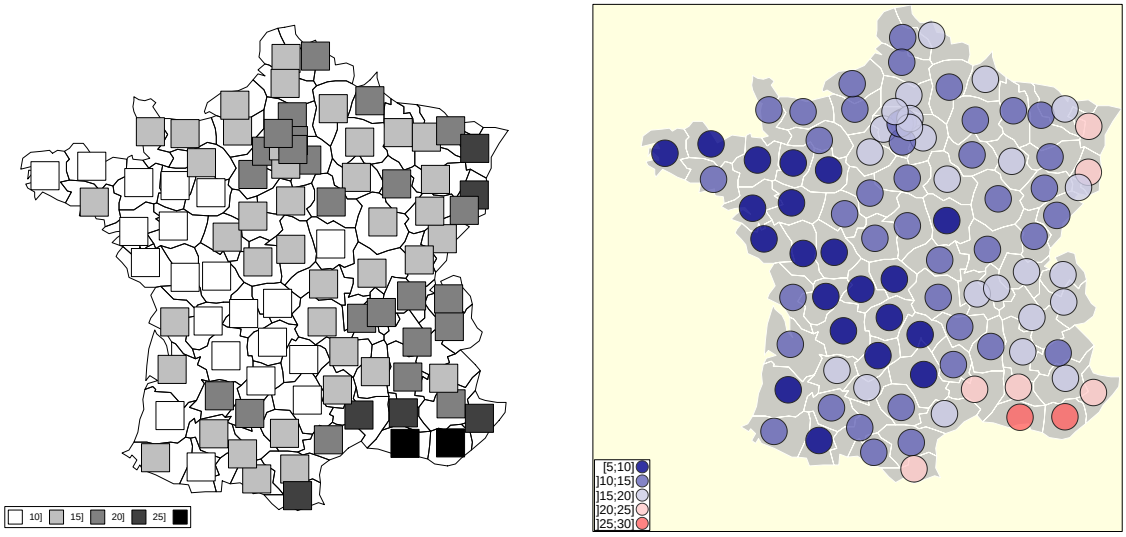
Mots clefs : Analyse multivariée, Graphique, Visualisation.

Le package `ade4` [1] propose de nombreuses fonctions utilisées aussi bien pour l'analyse exploratoire de données multivariées que pour leur représentation graphique. Si la partie procédurale a été enrichie grâce au développement de nouvelles méthodes d'analyse, les améliorations concernant les fonctionnalités graphiques ont été minimes. Au fil du temps, l'utilisation du package a donc révélé une faiblesse dans la flexibilité et l'adaptabilité des représentations graphiques et une difficulté à manier des graphes de plus en plus complexes.

Pour optimiser la prise en charge des graphiques liés à `ade4`, nous venons de développer le package `adegraphics`. Ce package réimplémente et améliore l'ensemble des fonctionnalités présentes dans `ade4` en s'appuyant sur l'environnement graphique fourni par le package `lattice` [2] et en adoptant une programmation orientée objet basée sur l'utilisation du format de classe S4. Il en résulte une amélioration de la représentation, de la gestion et de la manipulation des graphiques aussi bien pour les données brutes que pour les sorties d'analyse. Pour le développeur, l'utilisation d'une structure en classes facilite la maintenance du code et offre un environnement favorable à l'implémentation de nouvelles méthodes.

La structure adoptée dans `adegraphics` permet de stocker les graphiques sous la forme de deux grandes classes d'objets : graphique simple et graphique multiple. Cette dernière classe permet de gérer dans un seul type d'objet la juxtaposition, l'insertion et la superposition de graphiques. Un grand nombre de paramètres graphiques permet de modifier *a priori* et *a posteriori* les représentations associées à ces objets (Figure 1). Les structures complexes associées aux données multivariées nous ont conduit à proposer de nouvelles fonctionnalités permettant notamment de décliner facilement un même graphique en une collection via la prise en compte d'une partition des individus en groupe ou de multiples variables ou axes issus d'une analyse multivariée. L'identification de structures dans les données multivariées et leur restitution est ainsi simplifiée.

En adoptant une présentation basée sur l'analyse de jeux de données en écologie, nous présenterons les principaux atouts du package `adegraphics`. Nous explorerons notamment la représentation des données brutes, leur traitement, la restitution graphique des résultats d'analyse et son optimisation.



(a) (b)

Figure 1 - Exemple de l'évolution graphique avec le package `adegraphics` : cartographie des scores du premier axe d'une analyse en composantes principales. (a) Avec `ade4`, la représentation graphique est figée (le fond de carte, la forme et la couleur des symboles et la position de la légende sont fixés). (b) Avec `adegraphics`, les différents éléments du graphique (carte, légende, symboles) sont modifiables *a priori* et *a posteriori*.

Références

[1] Dray, S., Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**(4), 1-20

[2] Sarkar, D. (2008). Lattice: multivariate data visualization with R. *Springer Verlag*

Développement d'une application sous R pour la Surveillance Observationnelle des Problèmes de Santé au Travail

D. Rieutort[a] and **R. de Gaudemaris**[a,b] and **D.J. Bicout**[a,c]

[a] Environnement et Prédiction de la Santé des Populations (EPSP)
Laboratoire TIMC-IMAG, UMR CNRS 5525, Université Joseph Fourier
Domaine de la Merci, 38 706 La Tronche, France
Delphine.Rieutort@imag.fr

[b] Service de Médecine et Santé au Travail
Centre Hospitalier Universitaire Grenoble
B.P. 217, 38043 Grenoble, France
RDegaudemaris@chu-grenoble.fr

[c] Biomathématiques et Epidemiologie
EPSP-TIMC, UMR CNRS 5525, Université Joseph Fourier
VetAgro Sup, Campus Vétérinaire de Lyon, 1 avenue Bourgelat, 69280 Marcy l'Etoile, France
bicout@ill.fr

Mots clefs : Maladies professionnelles, surveillance observationnelle, modélisation, programmation, logiciel R.

Les travailleurs sont soumis dans l'exercice normal de leur activité professionnelle à des expositions d'origine et nature physiques, chimiques ou biologiques, qui peuvent avoir un impact sur le développement de pathologies. La surveillance de ces expositions et des maladies professionnelles associées est donc un enjeu de santé publique important, en particulier, pour identifier et prévenir les nouvelles menaces qui pourraient peser sur la santé des travailleurs. Dans ce contexte, le Réseau National de Vigilance et de Prévention des Pathologies professionnelles (RNV3P) a mis en place un réseau de médecins experts qui diagnostiquent et enregistrent, chaque année depuis 2001, dans une base de données tous les Problèmes de Santé au Travail (PST) [1] ; un PST étant défini par l'association d'une maladie ou pathologie (conséquence d'une exposition plus ou moins longue) et une exposition professionnelle composite, comprenant des nuisances potentiellement responsables et un contexte professionnel. Ce réseau a pour entre autres buts de développer des méthodes d'analyse et une expertise sur les relations maladies-expositions professionnelles.

Dans cet objectif, nous avons développé le concept d'exposome professionnel basé sur une utilisation optimale de la base RNV3P (environ 150 000 PST) et permettant d'investiguer les caractéristiques reliant ou séparant les PST [2]. A présent, nous sommes en train de développer la surveillance observationnelle des PST qui, basée sur l'exposome professionnel, consiste en la construction et la description des spectres dynamiques d'exposition d'une ou ensemble de pathologies. Cette surveillance observationnelle a pour but d'analyser et repérer les changements de tendance et détecter les évènements émergents dans les relations maladies-expositions professionnelles. Afin de pouvoir effectuer la surveillance observationnelle que nous développons de manière routinière et systématique, nous avons fait le choix du logiciel R pour implémenter nos méthodes d'analyses et traitements des données RNV3P ainsi que pour développer une

interface d'application.

Il reste encore certains points à développer, notamment l'introduction de l'aspect dynamique de nos résultats (par exemple les spectres), grâce au Graphics Interchange Format (GIF), également disponible dans R. Ensuite, nous souhaiterions produire pour chaque analyse effectuée, un rapport contenant les résultats principaux (actuellement exportés en fichiers tableurs et/ou images), afin de générer des fiches de surveillance des pathologies professionnelles. En conclusion, nous réalisons que l'utilisation du logiciel R pour notre problématique présente plusieurs fonctionnalités et avantages, notamment en permettant de faire à la fois la modélisation, la programmation et la construction de l'interface.

Références

[1] ANSES (2013). Le Réseau National de Vigilance et de Prévention des Pathologies Professionnelles. Disponible au : http ://www.afssa.fr/ET/PPN5BDA.htm ?pageid=1175&parentid=523 [*dernière consultation le 11/03/2013*]

[2] Faisandier, L., Bonneterre, V., De Gaudemaris, R., Bicout, DJ. (2011). Occupational exposome : a network-based approach for characterizing Occupational Health Problems. Journal of biomedical informatics, 44(4), 545-52

Estimation des données manquantes en morphométrie : quelle limite choisir ?

J. Clavel [a], G. Merceron [b], G. Escarguel [a]

[a] Laboratoire de Géologie de Lyon : Terre, Planètes, Environnements
UMR 5276 CNRS, ENS Lyon & Université Lyon 1
Campus de la Doua, 2 rue Raphaël Dubois, 69622 Villeurbanne.
Julien.clavel@univ-lyon1.fr, gilles.escarguel@univ-lyon1.fr

[b] Institut International de Paléoprimatologie Paléontologie Humaine : Evolution et Paléoenvironnements
UMR 7262 CNRS & Université de Poitiers
Bat. 8, 5 rue Albert Turpain, 86022
Gildas.merceron@univ-poitiers.fr

Mots clefs : Imputations multiples, données manquantes, morphométrie, simulations, ordinations, superimpositions procrustes.

Les estimations des dynamiques évolutives et de la diversité passée sont essentiellement basées sur l'étude de la variation morphologique de spécimens fossiles. Malheureusement, les restes fossiles sur lesquels de telles estimations doivent être effectuées sont souvent altérés par les processus post-mortem ou taphonomiques. Une telle perte d'information conduit souvent au retrait de certains spécimens dans les analyses multivariées et exclu de possibles comparaisons contrôlées statistiquement. Afin de contourner ce problème de données manquantes, des méthodes d'imputations sont souvent utilisées pour directement remplacer les valeurs manquantes par des estimations établies sur la partie non altérée du jeu de données. Cependant la proportion de valeurs manquantes dans un jeu de données peut conduire à des estimations significativement biaisées.

Ces dernières années, plusieurs seuils empiriques représentant la proportion maximale de données manquantes, que l'on peut considérer comme acceptable pour l'utilisation de techniques d'imputations, ont été proposés dans la littérature. D'un autre côté, certaines études ont critiqués ces seuils car ils sont souvent spécifiques aux jeux de données utilisés dans les simulations, à la distribution des valeurs manquantes, ou encore aux méthodes d'imputations utilisées, et ne sont donc en aucun cas généralisable.

Alternativement, des méthodes d'imputation multiples (MI) ont été développées pour considérer explicitement l'erreur associée aux estimations. Ces méthodes permettent

d’imputer *m* (>1) fois le même jeu de données via des processus de Monte Carlo. La variabilité obtenue sur ces *m* (>1) tableaux imputés, permet d’évaluer l’erreur associée aux estimations des valeurs manquantes.

Dans cette étude, nous évaluons les performances relatives de sept techniques d’imputations multiples disponibles sur R. Chacune des simulations ont étés effectuées sur un jeu de données morphométriques dégradé artificiellement suivant trois types de biais (aléatoires, anatomiques, et taxinomiques). Les simulations révèlent que les algorithmes FCS (Fully Conditional Specification) et EM (Expectation-Maximization) des packages MICE et Amelia II, produisent les meilleurs compromis statistiques en termes d’erreur systématique et de probabilité de recouvrement pour l’intervalle de confiance à 95%. De plus, les techniques d’imputations multiples apparaissent remarquablement robustes aux transgressions des conditions statistiques qui leur sont propres, comme par exemple la distribution non-aléatoire des données manquantes dans les jeux de données. Ces résultats montrent que les différences observées entre les types de distributions (aléatoire, taxinomiques, anatomiques) sont plus faibles qu’entre les méthodes d’imputations multiples elles-mêmes. Sur la base de ces résultats, plutôt que de proposer une valeur ou un ensemble de valeurs seuils, nous développons une approche qui combine l’utilisation de ces imputations multiples avec la super-imposition Procruste des résultats d’analyses en composantes principales. L’erreur associée à des individus pour lesquels certaines valeurs manquantes ont été imputées, peut être ainsi directement visualisée dans un espace ordonné.

Prédiction de la réactivité du glycérol sur les catalyseurs métalliques

Jérémie Zaffran*, Carine Michel, Françoise Delbecq, Philippe Sautet**

Laboratoire de Chimie, UMR ENSL-CNRS 5182, 46 allée d'Italie, 69364 Lyon Cedex 07
*jeremie.zaffran@ens-lyon.fr
**philippe.sautet@ens-lyon.fr

Mots clefs : régression linéaire, prédiction, erreurs, boite à moustaches, catalyse, DFT

Du fait de l'amenuisement des ressources pétrolières, la valorisation de la biomasse est un défi de taille pour les années à venir. Les réactions chimiques impliquées nécessitent souvent l'utilisation de catalyseurs hétérogènes (solides) tels que les métaux de transition. La modélisation moléculaire peut aider à la conception de catalyseurs de plus en plus performants.

Différentes techniques de la chimie théorique permettent d'étudier la réactivité d'un composé donné. Cependant les molécules issus de la biomasse sont souvent très complexes (près d'une quinzaine d'atomes pour les plus petites) et les techniques classiques de modélisation sont trop couteuses en temps de calculs pour être applicables. C'est pourquoi, à partir d'analyses statistiques réalisées avec le logiciel R [1] sur des calculs DFT effectués avec le code VASP, [2] nous avons établi des modèles pour prédire efficacement leur réactivité.

Nous nous intéressons dans ce projet à la déshydrogénation sur le rhodium d'un alcool complexe, le glycérol. Les modèles que nous avons établis reposent sur les relations de type BEP (Brønsted-Evans-Polanyi). Ces dernières visent à prédire une grandeur cinétique (l'énergie de l'état de transition par exemple) à partir d'un paramètre thermodynamique (l'énergie de l'état final par exemple), [3] les propriétés thermodynamiques étant usuellement plus faciles à calculer que les propriétés cinétiques (quelques heures contre plusieurs jours). (*fig1*)

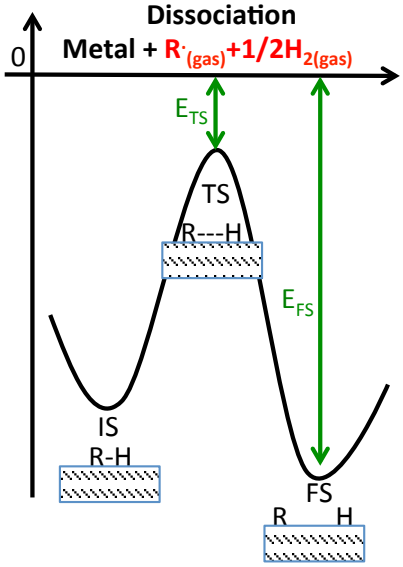


Figure 1: Exemple d'une relation de type BEP dans le cas d'une réaction de déshydrogénation d'une molécule R-H sur une surface. Ce type de relation exige de prendre une référence énergétique. IS : état initial, TS : état de transition, FS : état final

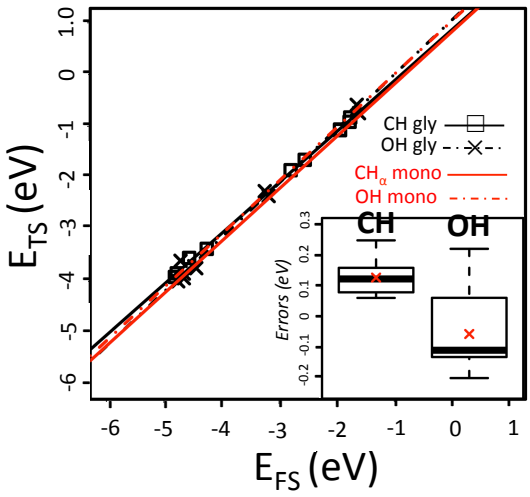


Figure 2: Prédiction de l'énergie des états de transition (E_{TS}) de la déshydrogénation du glycérol sur rhodium à partir de l'énergie des états finaux (E_{FS}), via le modèle obtenu pour les monoalcools. Les distributions d'erreurs pour les dissociations CH et OH sont représentés dans les boîtes à moustaches en bas à droite.

Nous avons considéré un échantillon de monoalcools se déshydrogénant sur le rhodium. Nous avons distingué les sous-ensembles « dissociations CH_α » et « dissociations OH ». Les régressions linéaires établies pour ces échantillons entre les énergies de leurs états de transition et de leurs produits, révèlent des erreurs moyennes absolues (MAE) inférieures à 0.10 eV. Nous avons ensuite appliqué ces modèles au glycérol, et nous avons observé que les énergies des états de transition du glycérol se déduisent de ses états déshydrogénés (produits) via les relations de type BEP établies pour les alcools simples avec des erreurs systématiques de +0.10 eV et -0.10eV, respectivement pour les dissociation CH et OH. (*fig2*)

Ce travail est donc utile pour réaliser un screening rapide des chemins réactionnels les plus favorables pour des molécules complexes sur un métal donné. Cette méthode appliquée de façon systématique sur différents métaux permet enfin de sélectionner le catalyseur optimal pour une réaction donné avec un gain de temps considérable. (*fig3*) En effet, une étude complète de réactivité pour des molécules relativement simples issues de la biomasse nécessite l'optimisation de plus d'une dizaine d'états de transition, soit plus d'un mois de calcul. L'application de cette méthode ramène ce temps à quelques jours seulement.

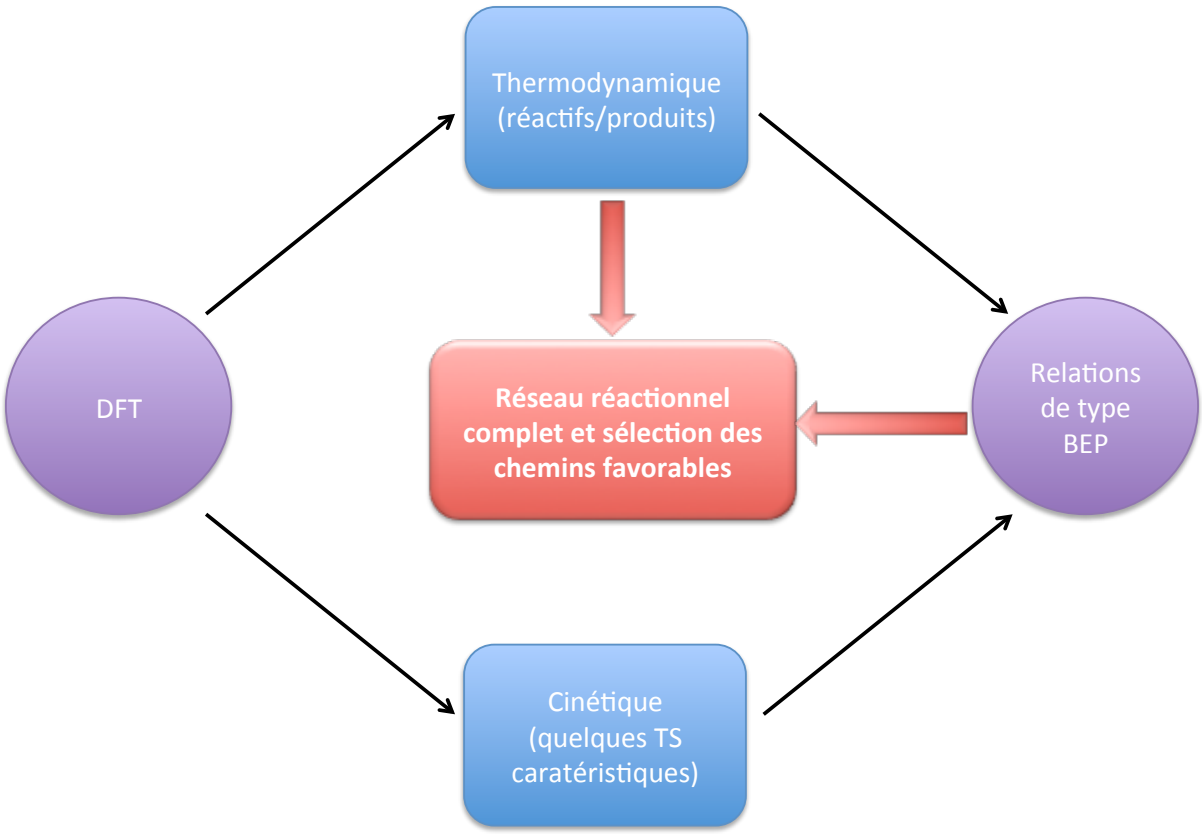


Figure 3: Schéma récapitulatif de l'obtention des relations de type BEP et de leur utilisation pour la conception in silico de catalyseurs plus performants.

Références

[1] http://www.r-project.org
[2] G.Kresse, J. Hafner, Phys. Rev. B 1993, 47, 558
[3] D. Loffreda, F. Delbecq, F. Vigné, P. Sautet, Angew. Chem., Int. Ed. 2009, 48, 8978-8980

What a statistician might want to know about human color vision, but was afraid to ask!

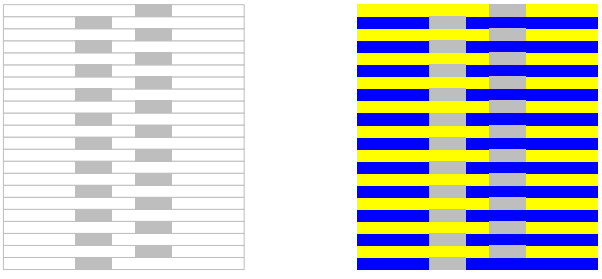
K. Knoblauch[a]

[a]INSERM, U846, Stem Cell and Brain Research Institute
Department of Integrative Neurosciences
18 avenue du Doyen Lépine, 69675 Bron
ken.knoblauch@inserm.fr

Mots clefs : visualisation, graphiques, couleur

The use of color can provide a powerful and effective means to enhance the salience of graphical displays. Misused, however, it can also mask the message that the author intends. Within the base package **grDevices** in R [5], there can be found several sophisticated facilities for manipulating the color of graphics. In addition, several add-on packages extend these capabilities. For those who might lack an artistic sense of how best to use color in data displays or for the roughly 8% of the population (mostly males) with abnormal color vision, it can be useful to understand some basic aspects of human color vision.

A number of confusions can be avoided by making a distinction between the concepts of *coding* and *appearance*. Coding concerns how the physical stimulus (e.g., light intensity, spectral distribution, etc.) is sampled by the visual system and transduced into neural activity; appearance concerns what a stimulus looks like. In this talk, I will highlight display issues with respect to these two concepts and some of the facilities available in R for dealing with them. While there is certainly a relation between the coding of light by the visual system and its color appearance, such a relation is complex and remains a research topic. As shown below, for example, lights that are encoded identically by the visual system can appear very differently as a function of context.



The two columns of small grey boxes in the left image appear quite different in the context of the alternating blue and yellow slats in the right image but are physically identical in both images (i.e., specified in each case by the argument `col = "grey"`*) and therefore encoded identically at the visual input.*

The trichromatic model of human color vision describes how chromatic differences are encoded in three spectrally different channels in the visual system. Color matching experiments that support the model have led to color specification systems for the spectral coding of lights by an average normal, observer (e.g., the CIE 1931 *xy* diagram) These systems, however, do *not* represent color appearance! For example, the grey boxes in the above figure have the same tristimulus values in the left and right images independently of the color of the adjacent slats.

Chromatic discrimination experiments demonstrate that the CIE *xy* diagram is anisotropic. This has led to the development of so-called uniform chromaticity diagrams (e.g., $L^*u^*v^*$ and $L^*a^*b^*$) that are, in fact, not uniform. The **colorspace** package facilitates specifying lights in R with respect to these different spaces [2]. Chromatic discrimination across the *xy* chromaticity diagram can be explained by a decorrelation of the signals at the retinal output [3].

It is estimated that about 1.3% of the population have dichromatic color vision, i.e., color matching behavior that depends on only two instead of three variables. The loss of one of the normal spectral channels leads to a collapse of the normal space to a two-dimensional subspace. There are three types of dichromacy (each corresponding to loss of one of the normal spectral channels). The **dichromat** package [4] provides color palettes that can be used to modify a graphic so that a color normal observer can appreciate for a given graphic display the loss of salience experienced for an observer with a particular type of dichromacy.

Recent work suggests how models of coding might be used to predict the salience of lights in a multicolor display. Salience can be gauged operationally by the reaction time to detect a particular color target in a field of distractor colors. Under some conditions, reaction time is independent of the number of distractors, suggesting high salience while under others, reaction time increases with the number of distractors, suggesting low salience with respect to the distractors. These results can be explained by models in which observers use a linear classifier to search for the target color [1].

What about appearance? The appearance of lights can be represented by coordinates along three perceptual dimensions that code opponent pairs of colors: red-green, yellow-blue, white-black. This opponent color coordinate system, however, is not related linearly to the spaces described above that encode identity between lights. Such appearance diagrams suggest useful color scales for graphics. Nevertheless, care must be taken in their use, as color appearance, unlike spectral coding, depends strongly on the spatial configuation of lights (as shown in the image above) and the adaptation state of the observer (e.g., for a transient effect of adaptation, stare at a fixed point on the right image above for 30 seconds and then move your gaze to the left image).

Références

[1] D'Zmura, M. (1991). Color in visual search. *Vision Research* **31**, 951–966.
[2] Ihaka, R., Murrell, P., Hornik, K., Fisher, J. C., Zeileis, A. (2013). colorspace: Color Space Manipulation. R package version 1.2-1. URL `http://CRAN.R-project.org/package=colorspace`.
[3] Le Grand, Y (1949) Les seuils différentiels de couleurs dans la théorie de Young. *Revue d'Optique*, **28**, 261–278.
[4] Lumley, T. (2013). dichromat: Color Schemes for Dichromats. R package version 2.0-0. `http://CRAN.R-project.org/package=dichromat`.
[5] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

KmL3D: K-means pour données longitudinales jointes

Christophe Genolini[1,2,⋆], **Jean-Baptiste Pingault**[3,4] **and Bruno Falissard**[4,5]

1. UMR 1027, INSERM, Université Paul Sabatier, Toulouse III, France
2. CeRSM (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La Défense, France
3. Research Unit on Children's Psychosocial Maladjustment, University of Montreal and Sainte-Justine Hospital, Montreal, Quebec, Canada
4. UI669 INSERM, Paris, France
5. University Paris-Sud and University Descartes, Paris, France
⋆Contact author: <christophe.genolini@u-paris10.fr>

Mots clefs : K-means, trajectoires jointes, partitionnement, interface graphique, graphes 3D dynamiques.

Les études longitudinales sont des études dans lesquelles les mêmes variables sont mesurées de manière répétée au cours du temps. Chaque suite de mesures, appelée variable-trajectoire, reflète l'évolution d'un phénomène. Ces études touchent des domaines variés (médecine, épidémiologie, économie, sociologie, informatique, physique,...) et sont de plus en plus nombreuses.
Ces dernières années ont vu se développer de nouveaux outils pour l'analyse de ces évolutions. Les plus largement utilisés sont les techniques de partitionnement (comme Proc Traj [1] ou KmL [2,3]). Elles consistent à grouper ensemble les individus dont les trajectoires se ressemblent et ainsi à définir des « trajectoires types » qui reflètent le comportement « moyen » des individus d'un même sous-groupe.
Les études longitudinales travaillent généralement non pas sur une mais sur de nombreuses variables-trajectoires. Se pose alors la question du partitionnement de plusieurs variables-trajectoires (appelées « trajectoires jointes »)

Si on note m le nombre de variables-trajectoires à analyser, la méthode d'analyse classique consiste à partitionner les m variables-trajectoires indépendamment les unes des autres, à obtenir ainsi m partitions P_i et de considérer comme partition finale la partition croisée $P = \prod_{1 \leq i \leq m} P_i$.

Figure 1: Graphe 3D dynamique. Cliquer sur le graphe avec le bouton gauche, puis faire bouger la souris pour changer de point de vue.

Or, de même que dans le cas classique les variables sont souvent corrélées, il est très probable que des variables-trajectoires évoluent conjointement. De plus, la richesse d'information contenue dans les trajectoires permet d'envisager des modes d'interaction bien plus complexes qu'une simple corrélation, ou qu'une monotonie conjointe. Malheureusement, la méthode des partitions croisées ne permet pas de détecter ce genre d'interactions complexes.
Une solution à ce problème consiste à partitionner simultanément les m trajectoires jointes. Pour cela, nous avons considéré un espace vectoriel de dimension $m+1$. Sur le premier axe, nous avons placé le temps. Chacun des m autres axes correspond à une variable-trajectoire. Nous avons ensuite défini une distance entre trajectoires jointes dans cet espace vectoriel. Au final, cela nous a permis d'appliquer k-means, un algorithme de partitionnement classique, aux trajectoires jointes. Un exemple en dimension 3 ($m = 2$ variables-trajectoires) est donné figure 1.
La procédure a été publiée dans [4], utilisée dans [5], puis programmée et mise à disposition de la communauté scientifique sous forme d'un package R, le package KmL3D [6]. Disponible sur le site du CRAN, il est dédié au partitionnement des trajectoires jointes. Comme le package KmL, il propose à l'utilisateur un certain nombre de solutions face aux problèmes posés par le partitionnement des données longitudinales. 12 méthodes d'imputation des manquantes sont proposées ; des exécutions multiples de k-means, en variant les conditions initiales et / ou le nombre de groupes considéré, sont gérés automatiquement ; une interface graphique conviviale et interactive permet à l'utilisateur de représenter graphiquement les partitions obtenues. Enfin, dans le cas de deux trajectoires jointes (représentation graphique dans $\mathbf{R}^3$), il est possible d'exporter des graphiques « 3D dynamiques » vers des fichiers pdf. Ces graphes dynamiques offrent à l'utilisateur la possibilité de changer de point de vue, permettant ainsi une meilleure visualisation de la troisième dimension (voir figure 1).

Références

[1] Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, 29(3), 374-393.
[2] Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, 25(2), 317-328.
[3] Genolini, C., & Falissard, B. (2011). KmL: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, 104(3), 112-121.
[4] Genolini, C., Pingault, J. B., Driss, T., Côté, S., Tremblay, R. E., Vitaro, F., ... & Falissard, B. (2012). KmL3D: A non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine*.
[5] Pingault, J. B., Côté, S. M., Galéra, C., Genolini, C., Falissard, B., Vitaro, F., & Tremblay, R. E. (2012). Childhood trajectories of inattention, hyperactivity and oppositional behaviors and prediction of substance abuse/dependence: a 15-year longitudinal population-based study. *Molecular Psychiatry*.
[6] Christophe Genolini (2012). kml3d: K-means for joint Longitudinal data. *R package version 2.1.2*. http://CRAN.R-project.org/package=kml3d

MFAMIX : une extension de l'analyse factorielle multiple pour des groupes de variables mixtes

A. Labenne[a], **M. Chavent**[b,c], **V. Kuentz-Simonet**[a], **T. Rambonilaza**[a] and **J. Saracco**[b,c]

[a]IRSTEA, UR ADBX
33612 Cestas Cedex, France
amaury.labenne@irstea.fr

[b]Univ. Bordeaux, IMB, UMR 5251
F-33400 Talence

[c]INRIA, CQFD
F-33400 Talence

Mots clefs : analyse factorielle multiple, méthode PCAMIX, analyse de la qualité de vie

1 Introduction

L'Analyse Factorielle Multiple (AFM) est une méthode de réduction de dimension qui permet de prendre en compte le fait que les individus sont décrits par des variables naturellement structurées en groupes ou thématiques. Initialement l'AFM a été mise en place pour l'analyse de variables quantitatives (Escofier et Pagès [1]). Elle a ensuite été élargie à l'analyse de groupes de variables qualitatives (Escofier et Pagès [2]) puis à l'étude d'un tableau de données que l'on qualifiera de "semi-mixte", où chaque groupe peut être soit de type quantitatif, soit de type qualitatif. Cette dernière extension de la méthode, proposée par Pagès [3], permet la réduction de dimension dans un contexte où les bases de données deviennent de plus en plus composites. A ce titre, nous sommes confrontés à une variété de données complexes dans les travaux relatifs à la construction d'indicateurs du développement durable, il faut entendre par là l'état de l'environnement, de l'économie, de la santé, des conditions sociales des individus comme des communautés. Pour cela, nous optons pour une approche en termes de qualité de vie : en effet, l'analyse et la mesure du bien être et de ses différentes composantes constituent un indicateur pertinent pour l'évaluation des états des sociétés. Face à la multitude de variables issues de thématiques différentes (environnement, social, économie, démographie, etc.) disponibles pour décrire la qualité de vie, les méthodes multi-tableaux telles l'AFM sont une réponse pertinente pour l'analyse de ces données structurées en groupes. Dans cette problématique, les variables au sein d'une même thématique ne sont pas homogènes, mais mixtes dans le sens où elles peuvent être quantitatives ou qualitatives. L'écriture actuelle de l'AFM et son implémentation dans le package R FactoMineR (Husson et al. [4]) ne permettant pas d'intégrer des thématiques mixtes dans l'analyse, nous proposons une extension de l'AFM qui permet l'analyse de groupes mixtes via l'utilisation de la méthode PCAMIX, voir par exemple Chavent et al. [5].

2 Rappels sur la méthode PCAMIX

La méthode PCAMIX présentée par exemple dans Chavent et al. [5] est une méthode d'analyse factorielle pour des données mixtes, c'est à dire un mélange de variables quantitatives et qua-

litatives. Elle est similaire à la méthode d'analyse factorielle de données mixtes d'Escofier [6] et Pagès [7] dans la manière dont sont recodées les variables qualitatives et quantitatives. Cependant, Chavent et al. [5] propose une formulation de PCAMIX à l'aide d'une décomposition en valeurs singulières (SVD) des données préalablement transformées. Cela permet d'obtenir directement les composantes principales, les "loadings" des variables quantitatives (corrélation avec les composantes principales), les rapports de corrélation entre les variables qualitatives et les composantes, ainsi que les coordonnées principales des modalités des variables qualitatives.

3 La méthode MFAMIX

Cette méthode se distingue d'une ACP ou d'une ACM globale appliquée à l'ensemble des données dans la mesure où elle permet de prendre en compte la structure en groupes de l'ensemble des variables. Pour cela, l'AFM applique une pondération particulière aux variables selon leur appartenance aux différentes thématiques. Ainsi l'influence des groupes est équilibrée dans la construction des composantes principales globales.
Le principe général de l'AFM mixte (appelée MFAMIX dans la suite) repose essentiellement sur deux étapes. Tout d'abord, on analyse chaque thématique prise séparément avec la méthode PCAMIX. On obtient ainsi la plus grande valeur propre correspondant à chaque sous-tableau. Puis, on applique PCAMIX sur l'ensemble de toutes les variables prises en commun où chaque variable est pondérée par l'inverse de la première valeur propre de la thématique dont elle est issue. Ainsi l'influence de chaque groupe est équilibrée dans la construction des composantes principales globales. Nous proposons une écriture sous forme de SVD pour MFAMIX. Les codes R sont disponibles auprès des auteurs et feront l'objet d'un package R.
L'application de la méthode via les codes R associés sera illustrée sur des données socio-économiques relatives à la qualité de vie d'un ensemble de communes du bassin versant Adour-Garonne.

Références

[1] Escofier B et Pagès J (1983), Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation des vins rouges du Val de Loire, *Revue de statistique appliquée*, 31(2) : 43-59.
[2] Escofier B et Pagès J (1998), *Analyses factorielles simples et multiples*, Dunod, 3^{e} ed.
[3] Pagès J (2002), Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes, *Revue de statistique appliquée*, 50(4) : 5-37.
[4] Husson F, Josse J, Lê S et Mazet J (2012). FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R, *R package version 1.20. http ://CRAN.R-project.org/package=FactoMineR.*
[5] Chavent M, Kuentz-Simonet V, Saracco J, (2012), Orthogonal rotation in PCAMIX, *Advances in Data Analysis and Classification*, 6 : 131-146.
[6] Escofier B (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des données*, 4(2) : 137-146.
[7] Pagès J (2004), Analyse factorielle de données mixtes, *Revue de statistique appliquée*, 52(4) : 93-111.

Méthodes de couplage de deux K-tableaux et collections de graphiques

J. Thioulouse[a], **A. Siberchicot**[a], **A. Julien-Laferrière**[a], **A.B. Dufour**[a], **S. Dray**[a]

[a]UMR CNRS 5558 - LBBE "Biométrie et Biologie évolutive"
UCB Lyon 1 - Bât. Grégor Mendel
43 bd du 11 novembre 1918
69622 VILLEURBANNE cedex
jean.thioulouse@univ-lyon1.fr

Mots clefs : Analyse de données, Écologie, ade4.

Le nouveau package `adegraphics` [1] est une ré-écriture des fonctions graphiques du package d'analyse de données multivariées `ade4`, utilisant des classes S4, et basée sur le package `lattice` [2]. Parmi les nombreux avantages de ce nouveau package [3], figure en particulier la possibilité, grâce à `lattice`, de gérer des collections de graphiques, par juxtaposition, insertion et superposition. Cette possibilité s'inspire directement des anciennes versions du logiciel **ADE4**, mais elle avait été laissée de côté lors de la ré-écriture sous forme de package **R**.
D'autre part, les méthodes d'analyse de données dites "multitableaux", et en particulier les méthodes de couplage de deux K-tableaux, conduisent naturellement à produire des collections de graphiques. On peut ainsi collectionner les graphiques des différents tableaux, et des différentes variables ou des différents individus d'un ou de plusieurs tableaux. On peut également collectionner les biplots en superposant les individus et les variables de chaque tableau. Dans le cas des méthodes de couplage de deux K-tableaux, on peut aussi vouloir juxtaposer et/ou superposer les graphiques correspondant aux deux tableaux de chaque paire.
Dans le domaine de l'analyse de données écologiques, les méthodes de couplage de deux K-tableaux sont particulièrement intéressantes car elles permettent par exemple d'analyser la stabilité des relations espèces-environnement. Dans ce cas, un des K-tableaux est constitué des mesures de variables environnementales répétées au cours du temps, et le second par des mesures d'abondance floro-faunistique, également répétées.
Après un bref rappel sur l'utilisation des méthodes K-tableaux et 2K-tableaux dans le package `ade4`, nous présentons plusieurs exemples de mise en oeuvre des fonctions du package `adegraphics` pour réaliser des collections de graphiques coordonnées et adaptées aux objectifs des méthodes d'analyse de K couples de tableaux (méthodes 2K-tableaux). Trois méthodes sont plus précisément détaillées: BGCOIA, STATICO, et COSTATIS. Elles sont toutes les trois disponibles dans le package `ade4`, et un article récent ([4]) propose une comparaison des trois, ainsi qu'une mise en oeuvre interactive par le biais d'un site Web permettant de reproduire les calculs et les représentations graphiques à l'aide des anciennes fonctions graphiques du package `ade4`: http://pbil.univ-lyon1.fr/SAOASOPET/.

- La BGCOIA (Between Group Co-Inertia Analysis, [5]) est une analyse de coinertie intergroupes. Chaque tableau est un groupe, les moyennes des variables par tableau sont calculées et arrangées en deux tableaux qui sont ensuite soumis à une analyse de coinertie.
- STATICO (STATIS et Coinertie, [6]) est une Analyse Triadique Partielle portant sur la série de tableaux de covariance croisées de chaque couple de tableau.
- COSTATIS (Coinertie et STATIS, [4]) est une analyse de coinertie des deux compromis obtenus par l'Analyse Triadique Partielle de chaque K-tableau pris séparément.

Les trois étapes des méthodes K-tableaux (interstructure, compromis, intrastructure) fournissent des coordonnées factorielles qui peuvent être utilisées pour faire diverses représentations graphiques. Dans le cas de l'intrastructure, on projette en général les lignes et les colonnes des tableaux initiaux dans l'analyse du compromis, ce qui fournit de nombreuses autres possibilités de représentations graphiques.

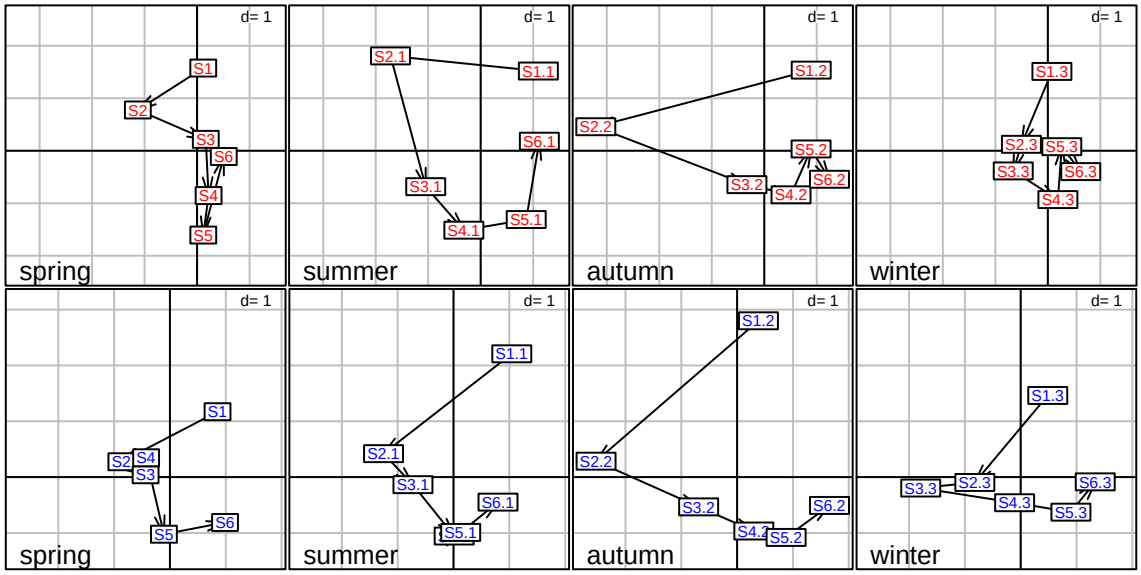


Figure 1: Intrastructure des stations pour les variables environnementales (en haut) et pour les espèces (en bas). Méthode STATICO, jeu de données `meau`, package `ade4`.

Par exemple, dans le cas de STATICO (Figure 1), on peut représenter la collection de graphiques obtenue par juxtaposition des graphiques des projections des individus par tableau dans l'analyse du compromis. L'utilisation du package `adegraphics` facilite alors grandement la réalisation de ces graphiques, en automatisant le processus de collection des graphiques par tableau, et en autorisant une grande souplesse dans leur positionnement et l'ajustement des paramètres secondaires (couleurs, type de points, labels, etc).

Références

[1] Julien-Laferrière, A., and Dray, S. (2012). Visualisation de données multivariées: réimplémentation des fonctionnalités graphiques de la librairie ade4. In Premières Rencontres R, Bordeaux, France.

[2] Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer, New York.

[3] Siberchicot, A., Julien-Laferrière, A., Thioulouse, J., and Dray, S. (2013) adegraphics : un package pour la représentation et l'analyse de données multivariées. In Deuxièmes Rencontres R, Lyon, France.

[4] Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *Annals of Applied Statistics*, **5**, 2300-2325.

[5] Franquet, E., Doledec S., and Chessel D. (1995) Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. *Hydrobiologia*, **300**, 425-431.

[6] Thioulouse J., Simier M. and Chessel D. (2004). Simultaneous analysis of a sequence of paired ecological tables. *Ecology*, **85**, 272-283.

ACP Fonctionnelles de Densités de Probabilité Estimées par la Méthode du Noyau Multivariée avec R

S. Yousfi[a], **R. Boumaza**[b] and **D. Aissani**[a]

[a]Laboratoire de Modélisation et d'optimisation des Systèmes
Université de Béjaia
Route de Targa Ouzemourt, 06000 Béjaia, Algérie
smail_yousfi@ymail.com
lamos_bejaia@hotmail.com

[b]Agrocampus ouest
INHP d'Angers
2 Rue André Le Notre 49045 Angers, France
Rachid.Boumaza@agrocampus-ouest.fr

Mots clefs : ACP Fonctionnelles, estimation à noyau, matrices de lissage.

L'ACP de densités de probabilité est une variante de l'analyse des données fonctionnelles bien adaptée aux données de type ternaires (instants × individus × variables). Ces données sont des tableaux $(n_t \times p)$ indicés par t, où à chaque instant t ($t \in \{1, \ldots, T\}$) on dispose d'un échantillon de taille $(n_t \times p)$ d'un vecteur aléatoire à p dimensions. En effet, en associant à chaque tableau t une densité de probabilité f_t (qu'on estime ensuite par la méthode du noyau multivariée), nous cherchons à apprécier globalement les différences et les ressemblances entre les tableaux via leurs densités associées. L'ACP sur ces T densités permet alors de faire ce travail à la façon dont procède la méthode STATIS dual dans sa deuxième étape (comparer globalement les T tableaux en se basant sur un tableau compromis, i.e., une combinaison linéaire des T densités), à différence de tenir compte dans les représentations aussi bien des moyennes que des variances-covariances des variables.

Le souci majeur lors de la mise en oeuvre de la méthode (ACP de densités) sur des données réelles réside dans le choix du critère de sélection des matrices fenêtres de lissage optimales utilisées dans l'estimation par noyau des densités. L'usage des critères de sélection classiques dans notre cas (Plug-in multivarié et Validation Croisée multivarié) engendre des temps de calculs trop important, qui sont proportionnels aux tailles des données, aux nombres de variables et aussi aux nombres des densités.

Dans cette communication nous proposons un critère de sélection de ces matrices basé sur le principe de maximisation du rapport inertie variance utilisé en ACP classique (Saporta 1990) et sur un choix a priori de ces matrices parmi une classe particulière de matrices de lissage (Fukunaga 1972). Les algorithmes permettant de réaliser ces calculs et ceux permettant de pratiquer l'ACP de densités ont été implémentés sous R et feront l'objet de package. La version d'essai est disponible auprès des auteurs et prochainement sur le site de CRAN. Notons que d'autre packages dédiés au vaste domain de l'analyse des données fonctionnelles existent dans la littérature ("*fda*", "*MFDA*", "*splines*" ...).

Nous illustrons sur des exemples de simulations (données issues d'un mélange de 3 gaussiennes bi-variée) le bon déroulement des algorithmes ainsi que le gain considérable en temps

d'exécution lorsque le choix est porté sur le critère de sélection (maximisation du rapport inertie variance).

Références

[1] Boumaza, R. (2004). Analyse en composantes principales de distributions gaussiennes multidimensionnelles. Revue de Statistique Appliquée, **XLVI** (2), 5-20.

[2] Duong, T. (2007). ks : Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R. *Journal of Statistical Software,* **21** (7).

[3] Kneip, A., Utikal, K.J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association,* **96** (454), 519-542.

[4] Lavit, C. (1988). Analyse conjointe de tableaux quantitatifs. *Masson*, Paris.

[5] Yousfi, S., Boumaza, R., Aissani, D. (2011). ACP de Densités de Probabilité et STATIS Dual. Estimation à noyau et choix de la fenêtre. *Actes des 43 ème Journées de la SFDS,* Gammarth, Tunisie

`aste`: An R package for the adaptive estimation the right tail

Frederico Caeiro[a]

[a]Faculdade de Ciências e Tecnologia & CMA
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
fac@fct.unl.pt

Keywords : Extreme value index, semi-parametric estimation, bias reduction.

Let us consider the common set-up of independent, identically distributed (i.i.d.) random variables (r.v.'s) $X_1, X_2, \cdots, X_n$, with a common distribution function (d.f.) F and denote the associated ascending order statistics (o.s.) by $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. Let us assume that there exist sequences of real constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that the maximum, linearly normalized, i.e., $(X_{n:n} - b_n)/a_n$ converges in distribution towards a non-degenerate r.v. Then F belongs to the max-domain of attraction of an *extreme value* (EV) d.f.,

$$EV_\gamma(x) = exp(-(1+\gamma x)^{-1/\gamma}), \quad 1+\gamma x > 0, \quad \gamma \in \mathbb{R}.$$

and we write $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$. The parameter γ is the *extreme value index* (EVI), the primary parameter of extreme events, with a low frequency, but a high impact. This index measures the heaviness of the right *tail function* $\overline{F} := 1 - F$, and the heavier the tail, the larger γ is.
The EVI needs to be estimated in a precise way, because such an estimation is one of the basis for the estimation of other parameters of extreme and large events, like a *high quantile* of probability $1-p$, with p small, the *right endpoint* of the model F underlying the data, $x^F := \sup\{x : F(x) < 1\}$, whenever finite, and the *return period* of a high level, among others. We will work with the $k+1$ top o.s.'s associated to the n available observations, assuming only that the model F underlying the data is in $\mathcal{D}_{\mathcal{M}}(G_\gamma)$. Most of the classical semi-parametric estimators of any parameter of extreme events have a strong bias for moderate up to large values of k, the number of top o.s.'s involved in the estimation, including the optimal k, in the sense of minimal mean squared error (MSE). Accommodation of bias of classical estimators of parameters of extreme events has been deeply considered in the recent literature. For the estimation of a negative or eventually zero EVI ($\gamma \leq 0$), we refer the recent *negative moment* estimator (Caeiro and Gomes, 2010),

$$\hat{\gamma}_{k,n}^{NM(\theta)} := \frac{1}{2}\left\{1 - \left(M_{k,n}^{(2)}/(M_{k,n}^{(1)})^2 - 1\right)^{-1}\right\} + \theta\, M_{k,n}^{(1)}, \quad \theta \in \mathbb{R}.$$

with

$$M_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\{\ln X_{n-i+1:n} - \ln X_{n-k:n}\}^j, \qquad j \geq 1, \qquad X_{n-k:n} > 0.$$

Apart from the usual integer parameter k, related with the number of top order statistics involved in the estimation, the estimator depend on an extra real parameter θ, which makes it flexible and possibly second-order unbiased for a large variety of models in $\mathcal{D}_{\mathcal{M}}(EV_\gamma)_{\gamma<0}$.
The package `aste` (adaptive short tail estimation) provides the *Algorithm* in Gomes *et al.* (2013) for the adaptive choice of the *tuning* parameters θ and k in the semi-parametric estimation of the EVI through such an estimator. The package also covers the estimation of high quantiles and the right endpoint of the model F underlying the data.

Example

As an example, we apply the *Algorithm* to the analysis of a set of environmental data, the daily average wind speeds in knots (one nautical mile per hour), collected in Dublin airport, in the period 1961-1978. Due to the seasonality of wind data, we restrict ourselves to the Autumn season data of size $n = 1602$. Spring and Summer data were already analyzed in Gomes *et al.* (2013).

Figure 1 (left) illustrates, for $\theta = 0,\ 1,\ 1.5$ and 2, the behaviour of $\hat{\gamma}_{k,n}^{NM(\theta)}$ as function of k. Notice that the parameter θ has a big influence on the sample path of $\hat{\gamma}_{k,n}^{NM(\theta)}$, as function of k. The application of the adaptive choice of θ, proposed in Gomes et al. (2013) led us to $\hat{\theta} = 1.604$. Figure 1 (right) shows the estimates $\hat{\gamma}_{k,n}^{NM(\theta)}$ and the 95% confidence limit, as function of k, with the adaptative $\hat{\theta} = 1.604$.

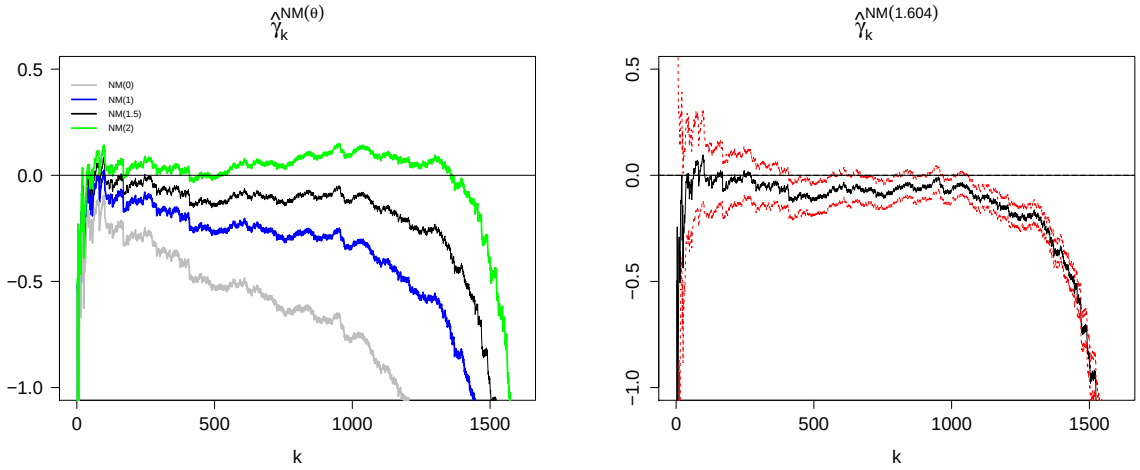


Figure 1: Left: Estimates of the EVI as function of k provided by the estimator under consideration, with several values of θ, for the Autumn wind speed dataset. Right: Estimates of the EVI (black) and 95% confidence limit (red), provided by the estimator under consideration with $\hat{\theta} = 1.604$.

The use of the Algorithm for the choice of k led us to $\hat{k} = 1048$ and to the adaptive EVI-estimate $\hat{\gamma} = -0.046$, and a 95% confidence interval $(-0.1045; 0.0129)$.

Acknowledgements. Research partially supported by National Funds through FCT— Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0297/2011 (CMA/UNL) and EXTREMA, PTDC/MAT/101736/2008.

References

[1] Caeiro, F. and Gomes, M.I. (2010). An asymptotically unbiased moment estimator of a negative extreme value index. Discussiones Mathematica: Probability and Statistics **30**(1), 5–19.

[2] Gomes, M.I., Henriques-Rodrigues, L. and Caeiro, F. (2013). Refined Estimation of a Light Tail: an Application to Environmental Data. Accepted in Torelli, N., Pesarin, F., Bar-Hen, A. (Eds.), Advances in Theoretical and Applied Statistics, Springer.

Cascade : un package R pour étudier la dispersion d'un signal dans un réseau de gènes.

F. Bertrand[a], **N. Jung**[a,b], **M. Maumy-Bertrand**[a], **S. Bahram**[b] and **L. Vallat**[b]

[a] Institut de Recherche en Mathématique Avancées (IRMA)
Laboratoire d'Excellence IRMIA
Université de Strasbourg, 67084 Strasbourg Cedex, France

[b] Laboratoire d'Immunogénétique Moléculaire Humaine,
Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche S1109
Laboratoire d'Excellence Transplantex
Université de Strasbourg, 67085 Strasbourg Cedex, France

Correspondance : njung@math.unistra.fr

Mots clefs : Statistique, Biologie, Lasso, Réseau de régulation génique.

Un réseau de régulation est un outil de modélisation de systèmes complexes particulièrement bien adapté pour étudier les interactions entre des gènes. En effet, certains gènes activés ont la capacité de moduler l'expression (ARN messager) d'autres gènes, formant ainsi un système complexe qui peut-être modélisé par un réseau (orienté ou non) dans lequel les nœuds correspondent aux gènes et les flèches correspondent à l'action d'un gène sur un autre.

Pour inférer le réseau de gènes, il est possible d'étudier le niveau d'expression de ces derniers grâce à des microarrays qui permettent de mesurer la quantité d'ARN messager produite par chaque gène activé. Afin de pouvoir déterminer un lien de causalité, il est important de mesurer l'expression des gènes au cours du temps. Ces réseaux peuvent alors être modélisés sous forme d'interactions en cascade [1] (Figure 1), mais peu d'outils ont été développés à ce jour pour appréhender ces phénomènes.

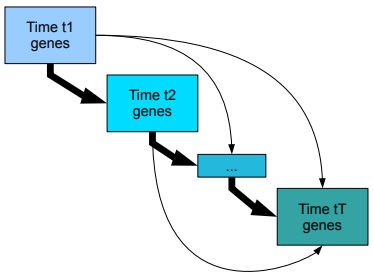


Figure 1: Réseau de gènes en cascade.

Le package R **Cascade** permet la sélection de gènes, l'inférence d'un tel réseau en cascade et la prédiction des effets d'une perturbation d'un ensemble de gènes dans le réseau (adapté de [1]). Une attention particulière a été portée dans la construction de ce package à la réalisation de graphiques facilement compréhensibles et interprétables par les biologistes. Les packages **animation** et **igraph** de R permettent ainsi de visualiser la propagation d'un signal dans le réseau. Vous pouvez trouver à cette adresse[1] un exemple.

[1] `http://www-irma.u-strasbg.fr/~njung/network/network.html`

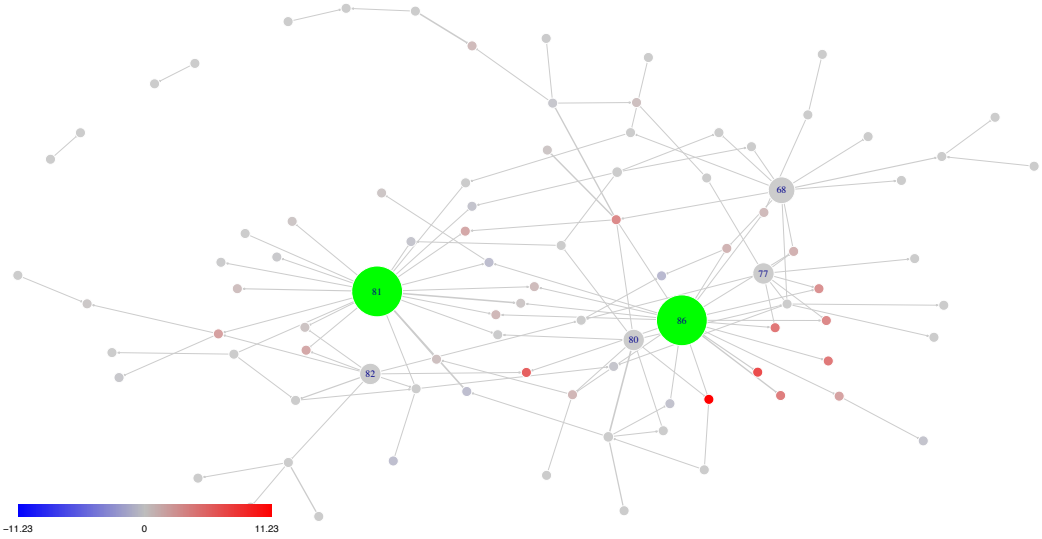


Figure 2: Prédiction de l'impact d'une intervention sur deux gènes du réseau (en vert).

L'étape de sélection des gènes permet de retenir des gènes fortement exprimés et des gènes présentant des cinétiques d'expressions spécifiques. Ceci permet notemment d'enrichir la sélection avec des gènes faiblement exprimés, mais présentant des cinétiques remarquables, en particulier aux temps précoces de la cascade [1]. Pour ce faire, nous avons construit des fonctions qui font une pleine utilisation des possibilités du package R de bioconductor **limma**.

Notre méthode d'inférence est basée sur celle développée dans [1]. C'est un modèle de régression linéaire pénalisé par une contrainte Lasso. De plus, la structure de réseau en cascade (Figure 1) permet d'assigner les gènes à des groupes temporels qui sont ensuite utilisés dans le modèle pour décrire l'action d'un gène sur un autre via les matrices $\boldsymbol{F}$:

$$\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{F}_{m(X_i)m(Y)} \omega_i \boldsymbol{X}_i + \lambda \sum_{i=1}^{N} |\omega_i| + \boldsymbol{\eta}, \tag{1}$$

où $\boldsymbol{Y}$ est le gène régulé et les $\boldsymbol{X}_i$ sont les gènes potentiellement régulateurs, les ω_i déterminent la puissance du lien entre X_i et $\boldsymbol{Y}$, et $m(.)$ est la fonction qui à un gène associe son groupe temporel ; λ est un cœfficient réel estimé par validation croisée déterminant la parcimonie du modèle et $\boldsymbol{\eta}$ est une erreur aléatoire. Plusieurs contraintes sont apportées afin d'assurer une évolution temporelle en cascade (voir [1] pour plus de détails).

A la fin du processus d'inférence, dans lequel chaque gène devient tour à tour gène régulé, un réseau en cascade est obtenu, matérialisé par les cœfficients ω. Cependant, cette méthode infère un grand nombre de lien. Or, il est parfois souhaitable pour le biologiste de réduire ce nombre de lien afin d'obtenir une information plus lisible. Pour cela nous proposons d'appliquer un seuillage sur les cœfficients ω. Notre package permet de voir l'évolution de la topologie du réseau en fonction de ce seuillage (voir exemple à cette adresse)[2]. En considérant la distribution du nombre de liens sortant (voir [2] par exemple), un test a été mis en place pour choisir le seuillage optimal. Des simulations ont par ailleurs confirmé l'intérêt d'un tel choix en montrant que cela permet de réduire le nombre de faux positifs parmi les liens inférés.

[2] `http://www-irma.u-strasbg.fr/~njung/evolution/evol.html`

Une fois le réseau inféré et le seuillage choisi, le résultat obtenu peut être visualisé sous plusieurs formes, dont la plus explicite est sans doute l'animation qui permet de voir un signal se propager dans le réseau [lien]. Pour finir, il est possible de prédire les effets d'une perturbation dans le réseau grâce au modèle donné dans l'équation (1), et de visualiser les changements prédits (voir Figure 2).

En conclusion, le package R **Cascade** a été conçu pour pouvoir être utilisé simplement et il apporte des représentations graphiques qui permettent une interprétation aisée des résultats. Il est disponible sur demande.

Références

[1]Vallat, L., Kemper, C. A., Jung, N., Maumy-Bertrand, M., Bertrand, F., Meyer, N., ... & Bahram, S. (2013). Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 110(2), 459-464.

[2]Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.

Corrélations entre maillages 3D au moyen du logiciel R application à l'imagerie par IRM de l'accident vasculaire cérébral

Anaïs Rouanet, **Carole Frindel**, **David Rousseau**

Université de Lyon, Laboratoire CREATIS ; CNRS, UMR5220 ; INSERM, U1044 ; Université Lyon 1 ; INSA-Lyon, 69621 Villeurbanne, France.

carole.frindel@creatis.insa-lyon.fr

Mots clefs : Imagerie, Neuroscience, Corrélation.

Nous présentons une analyse statistique, réalisée avec le logiciel R, sur des images obtenues par imagerie par résonance magnétique (IRM) dans le cadre d'un suivi longitudinal de patients ayant subi un accident vasculaire cérébral (AVC). L'analyse s'effectue sur une cohorte d'environ 400 patients suivis à 4 dates différentes : heure d'arrivée aux urgences, deux heures après, deux jours après puis un mois après. Les patients de cette cohorte sont tous atteints d'un AVC, où une lésion est observée dans les images IRM. Celle-ci est segmentée manuellement aux 4 dates sous forme d'images binaires comme visibles sur la Fig. 1A. Une problématique clinique majeure est la prédiction de l'évolution de cette lésion en fonction des paramètres hémodynamiques des tissus cérébraux estimés par imagerie de perfusion durant les premières heures, comme illustrés sur la Fig. 2.

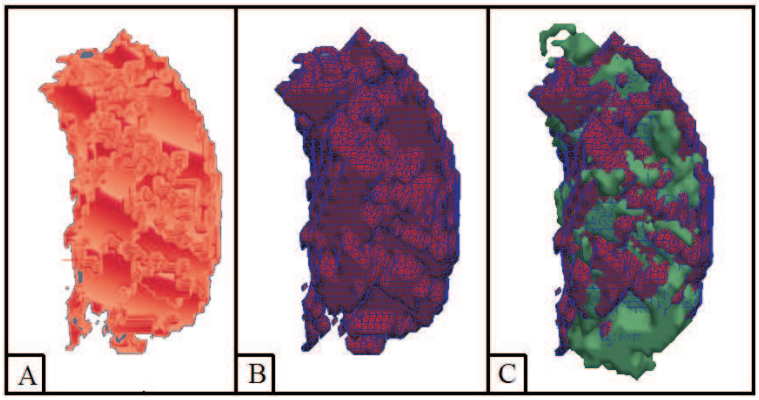


FIGURE 1 – Calcul des distances inter masques : A - Lésion au temps H0 (arrivée à l'hôpital). B - Maillage (marching cubes) de la lésion. C - Concaténation du maillage de la lésion à H0 et à H2 (2 heures après l'hospitalisation). Pour chaque sommet du maillage de la lésion à H0, la distance minimale au maillage de la lésion à H2 est calculée. On obtient alors une cartographie de distances sur le premier maillage.

L'approche classique actuelle consiste en une quantification globale de la corrélation entre des masques lésionels réalisés sur les différents paramètres hémodynamiques et le masque lésionnel final [1]. Cette approche ne tient pas compte du caractère non homogène des cartographies des paramètres hémodynamiques visibles sur la Fig. 2. Ce caractère non homogène est lié à l'hétérogénéité de vascularisation des tissus avec la présence de veines, d'artères de tailles variées. Nous proposons une approche locale pour quantifier la corrélation entre les masques lésionnels et les cartographies des paramètres hémodynamiques. Pour ce faire, nous réalisons un

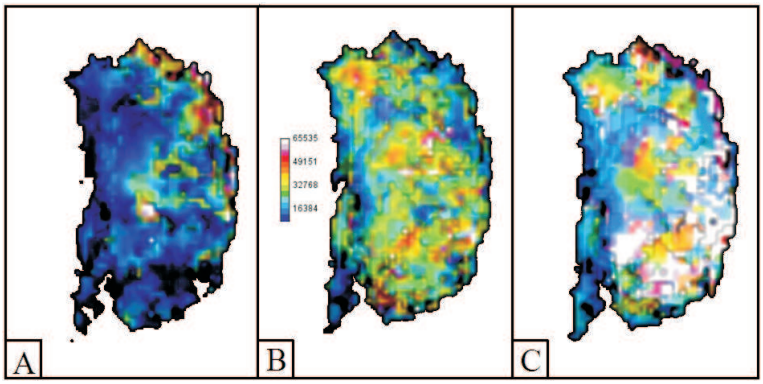


FIGURE 2 – Cartographies de 3 paramètres hémodynamiques caractéristiques de la perfusion cérébrale locale : A - Cerebral Blood Flow (CBF). B - Mean Time Transit (MTT). C - Time To Peak (TTP). Les cartographies sont appliquées au maillage de la lésion à H0.

maillage 3D des cartographies de lésion aux différentes dates et nous calculons, comme visible sur la Fig. 1, la distance entre ces deux maillages. Cette étape peut être réalisée au moyen de logiciels libres comme Paraview [2] pour le maillage et MeshValmet [3] pour la distance entre maillages. Les paramètres hémodynamiques sont ensuites appliqués sur le maillage de la première date comme visible sur la Fig. 2. Les données sont exportées sous forme de fichier texte. Nous calculons ensuite avec le logiciel R la corrélation entre le maillage de la Fig. 1 et ceux de la Fig. 2.

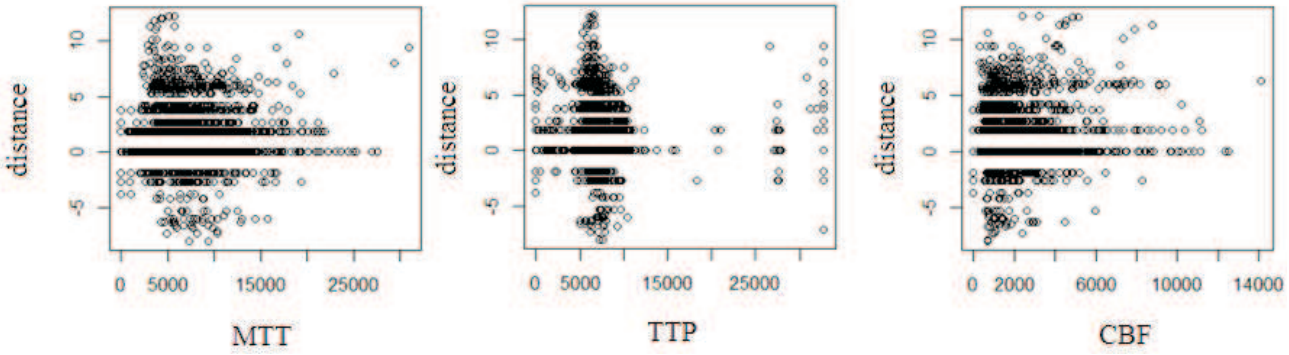


FIGURE 3 – Graphes réalisés sous le logiciel R de la distance en fonction de chaque paramètre hémodynamique pris séparément.

Nous observons sur la Fig. 3 le graphe de la distance entre maillages lésionnels aux deux premières dates et les différents paramètres hémodynamiques. Une analyse globale de la corrélation (méthode Pearson) sur ce patient donne une valeur de 0.036, ce qui n'est pas significatif. Il apparaît toutefois une forme de signature, comme visible sur la Fig. 3, puisque la dispersion des valeurs des paramètres hémodynamiques est moindre pour les points du maillage associés à une grande distance d'évolution (positive ou négative). Nous présenterons cette analyse plus en détails et sur de nombreux cas sur le poster de notre présentation.

Références

[1] Christensen, S. et al (2009). Comparison of 10 perfusion MRI parameters in 97 sub-6-hour stroke patients using voxel-based receiver operating characteristics analysis. *Stroke* **40**, 2055-61.
[2] http ://www.paraview.org/
[3] http ://www.nitrc.org/projects/meshvalmet/

Poster : Corrélations entre signaux EEG : un code d'analyse parallélisé

A. Cheylus [a], R. Fargier [a], and T. Nazir [a]

[a] Laboratoire sur le langage, le cerveau et la cognition (L2C2)
CNRS : UMR5304, Université Claude Bernard - Lyon I
Institut des Sciences Cognitives 67 bd Pinel 69675 Bron cedex - France
anne.cheylus@isc.cnrs.fr

Mots clefs : Statistique, EEG, Corrélation, Test de permutation, Calcul parallèle.

Les électroencéphalogrammes (EEG) de différents sujets ont été enregistrés lors de l'écoute de signaux auditifs, avant et après un apprentissage permettant d'apparier ces signaux à différentes modalités visuelles et motrices. Une modification des cross-corrélations entre les potentiels évoqués (ERP) associés à ces différents stimuli auditifs était recherchée avant et après apprentissage. Les détails de ce protocole seront donnés dans un article en préparation[1].

Pour analyser la significativité de différence de corrélation entre les ERP de différents sujets dans différentes conditions, une exploration par permutation des sessions avant et après apprentissage permet d'estimer la probabilité d'obtenir une telle différence de corrélation.

Afin d'explorer la totalité des permutations possibles, un code R a été développé qui parallélise l'exploration des corrélations obtenues pour chaque permutation et permet ainsi de bénéficier des différents processeurs du serveur de calcul pour obtenir plus rapidement le résultat.

Pour n sujets, il y a 2^n permutations de sessions à explorer. Si nous disposons de 2^p processeurs, chaque permutation possible des p premiers sujets sera attribuée à un processeur qui explorera les 2^{n-p} permutations possibles chez les n-p sujets restants.

Un appel à *save.image()* permet d'enregistrer l'espace de travail avant le lancement des calculs en parallèle. Les différences de cross-corrélation obtenues pour chaque permutation seront stockées dans des variables différentes pour chaque processeur. Le calcul parallèle est réalisé en lançant de nouvelles instances de R par appel à la fonction *system()*. Le système d'exploitation répartit ces instances entre les processeurs. Lorsque les calculs sont terminés, une nouvelle image est sauvegardée pour chaque processeur utilisé, puis un fichier est créé pour indiquer que l'image est prête à être lue. Les fonctions *load()* et *rbind()* sont utilisées pour agréger les résultats.

Références

[1] Nazir, T., Fargier, R., Cheylus, A., Paulignan, Y. and Reboul, A. (in prep.). Understanding each other through words: Inter-subject correlations of ERP signals during acquisition of novel words.

Parallel Computing in R using the BoT Package

Florent Chuffart

Laboratoire de Biologie Moléculaire de la Cellule
Ecole Normale Supérieure de Lyon
UMR5239 CNRS/ENS Lyon/UCBL/HCL
46, allée d'Italie
69364 Lyon cedex 07
florent.chuffart@ens-lyon.fr

Mots clefs : Distributed Computing, Bag-of-Tasks, Parameter Sweep.

BoT (stands for Bag Of Tasks) is an R package allowing to distribute independent tasks over many cores and many computing nodes. The simple fact that BoT is based on the process forking feature and task locking over file system makes BoT compatible with most of computing infrastructures: multicore, clusters, grids and clouds (see Figure 1).

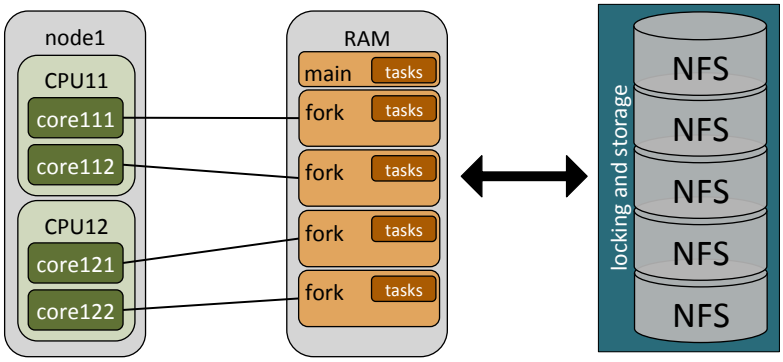


Figure 1: BoT architecture.

Using BoT, each task is a set of parameters associated with a user-defined function built on an R process. Next step consists in forking this R process for each core of the computing node. Finally, the forked set of tasks is randomized and executed in a parallel way. When a task starts a distributed lock is taken. This avoids redundant task execution. When a task is ended, result is dumped into a file.

As R package *mapReduce*, BoT uses a flexible parallelization backend. On the other hand BoT it isn't restricted to the MapReduce computational paradigm. Unlike R package *multicore* that is based on shared memory paradigm, BoT manages distributed memory: BoT is designed to be run on many heterogeneous computing ressources. BoT is based on the R package *fork*, it extends it in a fair way.

BoT is used to analyse ChIP-Seq data in the SiGHT project context (ERC-StG2011-281359). BoT has been used on two computing infrastructures: Grid'5000 experimental testbed [1] and PSMN computing center of ENS de Lyon [2]. BoT package is available on our web page[3].

[1] `https://www.grid5000.fr`
[2] `http://www.ens-lyon.fr/PSMN`
[3] `http://www.ens-lyon.fr/LBMC/gisv/index.php/en/protocols/bioinformatics`

Un *package* pour utiliser les *Cumulative Distribution Networks*

T. Pham[a] et **G. Mazo**[b]

Inria et Laboratoire Jean Kuntzmann
Inovallée, 655, av. de l'Europe
Montbonnot, 38334 Saint-Ismier cedex
[a]van-trung.pham@inria.fr,[b]gildas.mazo@inria.fr

Mots clefs : Cumulative Distribution Network, graphe, vraisemblance, fonction de répartition multivariée.

Un *Cumulative distribution network* (CDN) est une fonction de répartition d'un grand nombre de variables qui se factorise en produit de fonctions de répartition d'un plus petit nombre de variables (en pratique deux) et a été introduit par [1]. C'est donc un outil qui permet la construction de distributions en grande dimension [3]. On peut y associer un graphe ou les arrêtes représentent les fonctions reliant les variables. Prenons un exemple avec trois variables x_1, x_2, x_3 avec la structure de graphe représentée figure 1. Le CDN s'écrit alors

$$F(x_1, x_2, x_3, \theta) = \Phi_1(x_1, x_2, \theta)\Phi_2(x_2, x_3, \theta),$$

où Φ_1, Φ_2 sont deux fonctions de répartition choisies par l'utilisateur et θ est le vecteur des paramètres inconnus. Dans la pratique on choisit des fonctions de répartition paramétriques bivariées et se pose alors la question de calculer la vraisemblance $\partial_{x_1,x_2,x_3} F(x_1, x_2, x_3, \theta)$ et son gradient $\nabla_\theta \partial_{x_1,x_2,x_3} F(x_1, x_2, x_3, \theta)$ par rapport au vecteur des paramètres. Toujours dans [1], les auteurs ont proposé un algorithme de passage de messages qui permet leur calcul, ce qui autorise la maximisation de la vraisemblance en utilisant une méthode de type *quasi Newton* par exemple. Ce modèle a été utilisé par les auteurs pour modéliser un réseau de stations de pluviomètres et dans le cas d'un problème de *ranking* [2]. Toutefois l'implémentation délicate de l'algorithme peut freiner l'utilisateur dans la pratique. Nous nous proposons d'implémenter l'algorithme [1] et présentons ici un package permettant à l'utilisateur de modéliser ses données avec un CDN. Ce dernier pourra choisir la structure de graphe et les fonctions de lien dans différentes familles paramétriques et le calcul de la vraisemblance ainsi que du score lui sera retourné.

Références

[1] Huang, J.C., Jojic, N. (2010). Maximum-likelihood learning of cumulative distribution functions on graphs. *Journal of Machine Learning Research*, **9**, 342–349

[2] Huang, J.C. Cumulative distribution networks : Inference, estimation and applications of graphical models for cumulative distribution functions. *PhD thesis*, University of Toronto, 2009.

[3] Mazo, G., Forbes, F., Girard, S. Augmented cumulative distribution networks for multivariate extreme value modelling, *5th International Conference of the ERCIM WG on Computing and Statistics*, Oviedo, Espagne, 2012.

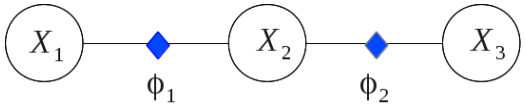


FIG. 1 – Exemple de structure en chaîne d'un CDN à trois variables.

Visualisation de processus spatiaux à l'aide de la correction de Ripley

A. Charpentier[a] and **E. Gallic**[b]

[a]Département de Mathématiques
UQAM-Université du Québec à Montréal
201, avenue du Président-Kennedy, Montréal (Québec), Canada H2X 3Y7
charpentier.arthur@uqam.ca

[b]Faculté des Sciences Économiques
Université de Rennes 1
7, Place Hoche – CS 86514, 35065, Rennes Cedex
ewen.gallic@etudiant.univ-rennes1.fr

Mots clefs : effets de bord, estimation par noyau, GIS, méthode de la circonférence de Ripley, polygones, processus spatiaux.

La connaissance de la date et de la position des accidents de voiture permet aux autorités publiques d'améliorer la sécurité routière. Dans la plupart des cas, on observe des regroupements d'accidents (*clusters*) dans des régions spatiales particulières, que l'on appelle également des *"points chauds"*. Dès lors, la mise en place de modèles spatiaux fournit une aide dans la détermination des endroits nécessitant une attention particulière.

L'analyse de la structure spatiale d'observations ponctuelles permet d'identifier ces *"points chauds"* ([1]). Une estimation de la densité peut être effectuée à l'aide de la méthode du noyau, où les pics représentent les *clusters* dans la distribution des événements. L'estimation fait intervenir une fenêtre (*bandwidth*) relative à la taille du voisinage des points considérés, ainsi qu'une fonction de pondération, le noyau. Parmi les différents noyaux qu'il est possible d'utiliser, le noyau gaussien semble être le plus populaire, grâce sa relative facilité de mise en œuvre, et ses bonnes propriétés statistiques. Cependant, l'emploi du noyau gaussien souffre d'un effet de frontière ([2]), qu'il est possible de corriger à l'aide de la méthode de la circonférence de Ripley, même si cette méthode était jusqu'à présent théorique car difficile à mettre en œuvre pour avoir des gains de calculs significatifs

L'objectif est de proposer une mise en œuvre facilement reproductible sur le logiciel R d'estimation de la densité en corrigeant des effets de bord. L'idée est de considérer des points $\boldsymbol{Z}_i$ appartenant à une surface $\mathcal{S}$ et de pondérer les observations par l'aire occupée par la fenêtre utilisée dans cette surface $\mathcal{S}$ (avec un lien simple entre la fenêtre de lissage et le rayon de la méthode de Ripley). Un exemple sur des données d'accidents de la route dans le département du Finistère est présenté pour illustrer la méthode. Une repésentation graphique des résultats est proposée sur une carte réalisée à l'aide des *packages* maps et ggmaps.

References

[1] Ripley, B. 1981. *Spatial Statistics*, Wiley, New York.
[2] Yamada, I. and Rogerson, P.A. 2003. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35: 97–109.

Visualisation et cartographie des données de capteurs météorologiques à l'échelle des terroirs viticoles

M. Madelin[a] and C. Bonnefoy[b]

[a] Université Paris Diderot - Sorbonne Paris-Cité, UMR 8586 PRODIG CNRS
5 rue Thomas Mann 75205 PARIS CEDEX 13, France
malika.madelin@univ-paris-diderot.fr

[b] Université Paris Diderot - Sorbonne Paris-Cité, UMR 8586 PRODIG CNRS
& Laboratoire COSTEL, UMR6554 LETG CNRS, Université Rennes 2 – Haute Bretagne,
35043 RENNES, France
cyril.bonnefoy@uhb.fr

Mots clefs : web mapping, climatologie, vignoble

Les répercussions du changement climatique s'observent dans de nombreux vignobles dans le monde : précocité des dates phénologiques, modification des conditions de maturation, etc. Les viticulteurs et la profession viticole sont alors demandeurs d'une connaissance des relations plante/environnement à une échelle fine, afin d'assurer la production de vins de terroirs de qualité, uniques et compétitifs sur le marché international. C'est dans cette optique que s'inscrit le projet GICC-TERADCLIM, qui s'intéresse à la variabilité climatique et à l'adaptation au changement climatique à l'échelle des terroirs viticoles : acquisition des données météorologiques et agronomiques sur plusieurs vignobles expérimentaux ; modélisation climatique (atmosphérique et statistique) ; intégration des scénarios du GIEC ; scénarios d'adaptation à une échelle de temps de 15-30 ans avec l'utilisation d'une plateforme multi-agents.
Dans le cadre de la première partie de ce projet, l'objet de cette communication, sous forme de poster, est de présenter la visualisation des données météorologiques acquises, sur une interface web, *via* le logiciel R installé sur un serveur. Les viticulteurs associés et les autres membres du projet peuvent ainsi interroger la base de données, choisir le paramètre étudié (température minimale, maximale, moyenne, …), la période et le pas de temps et visualiser en sortie le graphique et la carte des résultats (capteurs). Dans cette communication, il s'agit de décrire les principaux choix techniques de ce projet en cours et de discuter des avantages et inconvénients de cette solution.

R and the Cloud

K. Chine [a]

[a] Cloud Era Ltd
24 Langham Road,
Cambridge
UK
karim.chine@cloudera.co.uk

Mots clefs: Cloud Computing, EC2, e-Learning, distant education, e-Science, Collaboration, Science Gateways, Big data, Analytics-as-a-Service

Cloud Computing is holding the promise of democratizing access to computing infrastructures and deeply impacting research and education. However, the question "How will we bring the Infrastructure-as-a-Service paradigm to the data scientist's desk and to the statistics classroom?" has remained unanswered. The Elastic-R Software platform proposes new concepts and frameworks to address this question: R, Python, Matlab, Spreadsheets, etc. are made accessible as articulated, programmable and collaborative components within a virtual and immersive environment for scientific research and higher education.

Teachers can easily and autonomously prepare interactive R-based custom learning environments and share them like documents in Google Docs. They can use them in the classroom or remotely in a distant learning context. They can also associate them with on-line-courses. Students are granted seamless access to pre-prepared, controlled and traceable learning environments. They can share their R sessions to receive guidance from Teachers or to solve problems in collaboration. Costs may be hidden to the students by allowing them to access temporarily shared institution-owned resources or using tokens that a teacher can generate using institutional cloud accounts.

Scientists can easily use the cloud as a ubiquitous and scriptable collaborative environment for traceable and reproducible data analysis and computational research. The cloud becomes a user friendly Google-Docs-like platform where all the artifacts of computing can be produced by any number of geographically distributed real-time collaborators and can be stored, published and reused. Big data access and analysis are simplified and made accessible to wider range of research professional. Science Gateways (graphical user interfaces for data science; set of tools, applications, and data integrated via portals) are made R-scriptable and hence easy to create, publish and update on the fly from the R command line: their use becomes an intrinsic part of the process of programming with Data.

The presentation will give an overview of the synergies that exist between R and the state-of-the art cloud technologies. Elastic-R on Amazon's public cloud will be demonstrated via real-world applications in education, in bioinformatics and in finance.

Références

[1] Karim Chine (2010). Learning math and statistics on the cloud, towards an EC2-based Google Docs-like portal for teaching / learning collaboratively with R and Scilab, icalt, pp.752-753, 2010 10th IEEE International Conference on Advanced Learning Technologies.

[2] Karim Chine (2010). Open science in the cloud: towards a universal platform for scientific and statistical computing. In: Furht B, Escalante A (eds) Handbook of cloud computing, Springer, USA, pp 453–474. ISBN 978-1-4419-6524-0

[3] www.elastic-r.net

[4] aws.amazon.com

[5] www.coursera.org

[6] sciencegateways.org

R2GUESS: a GPU-based R package for sparse Bayesian variable selection

L. Bottolo[b], **M. Chadeau-Hyam**[b], **B. Liquet**[a], **S. Richardson**[a] and **H. Saadi**[b]

[a]MRC Biostatistics Unit
Institute of Public Health
University Forvie Site, Cambridge, UK
benoit.liquet@isped.u-bordeaux2.fr

[b]Department of Epidemiology and Biostatistics
Imperial College London
St Mary's Campus, London, UK
h.saadi@imperial.ac.uk

Mots clefs : Biology, Genomics, Bayesian, Variable Selection, GPU, High Dimensional Statistics.

Recent advances in high throughput "omics" technologies have given rise to a wealth of novel high dimensional data, ranging from thousands to hundreds of thousand variables, each demonstrating complex correlation structures. These data comprise genetic, epigenetic and transcriptomic profile which have shown a great potential in measuring the abundance of biologically relevant molecules over a whole biological system. The analysis of such complex data raises serious statistical challenges relating to the fact that the number of predictors exceeds the number of observations ("large p, small n" scenario).

Alongside multiple testing correction strategies, variable selection approaches are well suited to handle this situation, and we propose here a Bayesian implementation of this kind of approaches. As such, the method seeks for the best combination of covariates to predict the (possibly multivariate) outcome. The Bayesian framework it is based on allows for the construction of parsimonious regression models, adopting prior specifications that translate expected sparsity of the underlying biology, and therefore facilitating results interpretation.

R2GUESS is an R package that interfaces a C++ implementation of a fully Bayesian Variable Selection approach for multivariate linear regression . Using latest computational advancement, it can run on GPU (Graphical Processing Unit), and in its current form it enables the analysis of hundreds of thousands of predictors measured in thousands of individuals simultaneously. The efficient exploration of the 2^n dimensional space is possible thanks to the use of an Evolutionary Monte Carlo sampling scheme comprising a large portfolio of local and global moves. R2GUESS also provides refined numerical and graphical output facilitating post-processing and subsequent interpretation of the extensive output produced by the GUESS algorithm. Performances of the model and interpretability of its results are illustrated with examples from several omics platforms.

References

[1] Bottolo, L., Richardson, S., (2010). Evolutionary Stochastic Search for Bayesian Model Exploration. *Bayesian Analysis*, **5**, 583-618.
[2] Bottolo, L., Chadeau-Hyam, M., Hastie D.I., Langley, S.R., Petretto, E., Tiret, L., Tregouet, D., Richardson, S. (2011). ESS++: a C++ objected-oriented algorithm for Bayesian stochastic

search model exploration. *Bioinformatics*, **27**, 587-588.

An R package using HPC for entropy estimation and MCMC evaluation

D. Chauveau[a] and **P. Vandekerkhove**[b]

[a]MAPMO - Fédération Denis Poisson
Université d'Orléans et CNRS UMR 7349
BP 6759, 45067 Orléans cedex 2
didier.chauveau@univ-orleans.fr

[b]LAMA - CNRS UMR 8050
Université de Marne-la-Vallée
77454 Marne-la-Vallée cedex 2
Pierre.Vandekerkhove@univ-mlv.fr

Mots clefs : Entropy estimation, High Performance Computing, Kullback divergence, MCMC algorithms, Nonparametric statistics, Rmpi library.

Many recent (including adaptive) MCMC methods are associated in practice to unknown rates of convergence, leading to difficulties in assessing performance of specific MCMC samplers. Comparison or evaluation of MCMC samplers is now a challenge addressed by various approaches (see, e.g., the recent SamplerCompare package [6]). Let f be a d-dimensional target density of a MCMC algorithm, and p^t the marginal density of the algorithm at "time" (iteration) t. In Chauveau and Vandekerkhove [1], we have first proposed to evaluate a MCMC sampler efficiency from a Kullback divergence criterion,

$$\mathcal{K}(p^t, f) = \int p^t \log\left(\frac{p^t}{f}\right) = \mathcal{H}(p^t) - \mathbb{E}_{p^t}[\log f].$$

where $\mathcal{H}(p) := \mathbb{E}_p[\log p] = \int p \log p$ is the differential entropy of a probability density p over $\mathbb{R}^d$. We have introduced a simulation-based methodology allowing to estimate the entropy of the algorithm successive densities, $\mathcal{H}(p^t)$, based on the "parallel " simulation of N iid copies of (eventually Markov) chains at step t, resulting in a N-sample $\boldsymbol{X}^t$ iid$\sim p^t$, for $t \geq 1$. These simulations are first used to estimate $\mathbb{E}_{p^t}[\log f]$ (or more generally an estimate $\propto \mathbb{E}_{p^t}[\log f]$ if the normalizing constant of f is unknown) via standard Monte Carlo integration. The sample $\boldsymbol{X}^t$ is also used to compute an estimate of $\mathcal{H}(p^t)$, and we have proved in [1] some consistency results in this MCMC context for an entropy estimate based on Monte-Carlo integration of a kernel density estimate introduced by Györfi and Van Der Meulen [3]. Unfortunately, this estimate deteriorates as dimension increase, and require some parameters (like, e.g., the kernel bandwidth) whose tuning is challenging in practice.
We investigate here an alternative strategy based on Nearest Neighbor (NN) estimates of differential entropy, initiated by Kozachenko and Leonenko [4],

$$\hat{\mathcal{H}}_N(p^t) = \frac{1}{N}\sum_{i=1}^{N} \log(\rho_i^d) + \log(N-1) + \log(C_1(d)) + C_E, \tag{1}$$

where $C_E = -\int_0^\infty e^{-t}\log t\, dt$ is the Euler constant, $C_1(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ and where ρ_i is the nearest neighbor (Euclidean) distance from the ith point to the other points in the sample $\boldsymbol{X}^t$.

Kozachenko and Leonenko [4] proved, under mild conditions a mean square consistency of $\hat{\mathcal{H}}_N(p^t)$ for any dimension d. However, apparently, this NN approach has been used and studied mostly in univariate or bivariate ($d = 2$) situations (e.g., in image processing). We show that, in MCMC setup where moderate to large dimensions are common, this estimate seems more promising than kernel density estimates, both from an operational point of view (no tuning parameters like the bandwidth), and from a computational point of view (the nearest distance can be computed faster than a multivariate kernel density estimate in high dimension). Entropy estimation is also considered in other fields, and recent researchs extend the NN idea to a k-th nearest distance estimate (see, e.g., Singh et al. [5]), that we plan to investigate as well.
The computational burden required by our method can be heavy (depending on the dimension, kernel complexity, number of iid chains). We thus implement all our algorithms (iid MCMC simulation plus entropy and Kullback estimation) in the R package EntropyMCMC, which takes advantage of recent advances of High Performance Computing in R. This package can use MCMC output from samplers and target distributions implemented in other packages, such as, e.g., SamplerCompare [6]. The end user can also run its own MCMC inside the package by just providing R definitions for its target and, e.g., the proposal for standard Hastings-Metropolis or Independence samplers. Several functions are written with appropriate C and R code for running it on multicore computers, network of workstations or actual clusters, using e.g., the Rmpi library. We illustrate its usage for studying the behavior of the NN estimate in MCMC setup and moderate to large dimensions, using the cluster *Centre de Calcul Scientifique en région Centre* (`http://cascimodot.fdpoisson.fr/?q=ccsc`).

References

[1] Chauveau, D. and Vandekerkhove, P. (2012). Smoothness of Metropolis-Hastings algorithm and application to entropy estimation. *ESAIM: Probability and Statistics*.

[2] et Modélisation Orléans Tours, C. S. (2011). Centre de calcul scientifique en région centre.

[3] Györfi, L. and Van Der Meulen, E. C. (1989). An entropy estimate based on a kernel density estimation. *Colloquia Mathematica societatis János Bolyai 57, Limit Theorems in Probability and Statistics Pécs*, pages 229–240.

[4] Kozachenko, L. and Leonenko, N. N. (1987). Sample estimate of entropy of a random vector. *Problems of Information Transmission*, 23:95–101.

[5] Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003). Nearest neighbor estimate of entropy. *American Journal of Mathematical and Management Sciences*, 23(3):301–321.

[6] Thompson, M. (2010). *SamplerCompare: A framework for comparing the performance of MCMC samplers*. R package version 1.0.1.

Intégration R et C++ avec Rcpp

R. François[a]

[a]R Enthusiasts
1, place de l'égalité
42 400, Saint Chamond
author1@institut.com

Mots clefs : Rcpp, C++.

`R` est un langage parfaitement adapté à l'analyse statistique. `R` est un langage flexible et largement adopté par la communauté de statisticiens.

Cependant, de part sa nature de langage interprété, R peut ne pas fournir les performances suffisantes pour certaines applications, et il est parfois nécessaire de convertir des parties critiques du code dans un langage compilé comme `C++`.

`R` est implémenté en `C` et permet d'étendre ses fonctionnalités avec du code en `C` ou `C++`. L'API proposée par `R` est rudimentaire — faite de macros et fonctions `C` — et devient rapidement lourde à l'utilisation.

`Rcpp` fournit un jeu de classes permettant une intégration beaucoup plus simple d'utilisation. Avec plus d'une centaine de package dépendant, `Rcpp` est devenu un moyen populaire d'étendre les fonctionnalités de `R` avec du code performant.

Cette présentation propose un tour rapide de l'API `Rcpp`.

Références

[1] Eddelbuettel, D., François, R. (2011). Rcpp: Seamless R and C++ integration. Journal of Statistical Software.

[2] Eddelbuettel, D. (2013). Seamless R and C++ integration with Rcpp. Springer, useR! series.

`frailtypack` : Un package pour l'analyse de données de survie corrélées

A. Laurent[a] et **Y. Mazroui**[b] et **A. Mauguen**[a] et **V. Rondeau**[a,c]

[a]INSERM U897
ISPED
146 Rue Léo Saignat 33076 Bordeaux Cedex
Alexandre.Laurent@isped.u-bordeaux2.fr
Yassin.Mazroui@isped.u-bordeaux2.fr
Audrey.Mauguen@isped.u-bordeaux2.fr
Virginie.Rondeau@isped.u-bordeaux2.fr

[c]Université Bordeaux Segalen, ISPED
146 Rue Léo Saignat 33076 Bordeaux Cedex

Mots clefs : Survie, Données groupées, Modèles à fragilité, Modèles conjoints, Événements récurrents

Les modèles à fragilité sont une extension des modèles à risques proportionnels de Cox qui sont les plus populaires dans l'analyse de survie. Dans la plupart des applications cliniques, la population d'étude est considérée comme un échantillon hétérogène ou autrement dit un échantillon de plusieurs groupes homogènes pouvant représenter des familles ou des zones géographiques. Parfois pour des raisons économiques ou par omission, certaines variables étroitement liées à l'événement d'intérêt peuvent ne pas être mesurées. Les modèles à fragilité, ou modèles de survie à effet aléatoire permettent alors de tenir compte de l'éventuelle hétérogénéité de la population liée à des variables non mesurées.

`frailtypack` [1,2] est un package R qui proposent plusieurs types de modèles à fragilité pour données censurées à droite et tronquées à gauche. Le modèle à fragilité partagée (shared) qui est le modèle de base, est à utiliser lorsque nous savons que les données ne sont pas indépendantes. D'autre part, le risque de récidives peut être interrompu par le décès, qui peut être décrit comme une censure informative. Ainsi, il est possible de modéliser conjointement les fonctions de risque associées aux événements récurrents et terminaux en considérant un effet aléatoire commun aux observations d'un même sujet [3]. Dans la continuité de ce même modèle, nous pouvons obtenir un modèle conjoint pour des données groupées [4] où l'effet aléatoire serait commun aux individus d'un même groupe. Toujours dans l'idée de la modélisation conjointe, un modèle multivarié [5] a été implémenté et permet l'étude cette fois de deux types d'événements récurrents en plus d'un événement terminal.

L'estimation des fonctions de risque de base se fait toujours de manière non paramétrique en approximant par des splines ou plus récemment en utilisant une paramétrisation Weibull ou une approximation en constant par morceaux. Les modèles shared et joint qui sont les plus utilisés sont les premiers à bénéficier d'améliorations notables. Jusqu'à maintenant le package ne proposait qu'une distribution de type Gamma pour les effets aléatoires, il est maintenant possible d'y considérer une distribution Log-Normale. De plus, les variables explicatives sont supposées avoir un effet constant sur le temps d'apparition de l'événement d'intérêt, il est désormais possible de considérer des variables qui ont un effet dépendant du temps en modélisant les coefficients de régression par une combinaison linéaire de B-splines. Le package propose

aussi l'inclusion de variables dépendantes du temps, de stratification mais aussi de résultats de prédiction.

Le package dont la première version date de 2005 et qui proposait uniquement de faire un simple modèle à fragilité partagée a bien été enrichi depuis. Initialement, le programme était écrit en Fortran 77 jusqu'à son adaptation au logiciel R dont l'utilisation dépend et se veut très proche du package `survival`. Cependant encore aujourd'hui, les corps de programme sont toujours implémentés en Fortran 90. L'outil étant en constante évolution, l'objectif de cette présentation sera de mettre en lumière les récentes améliorations faites sur les modèles à fragilité partagée et les modèles conjoints et montrer leur utilisation dans un cas concret d'épidémiologie.

Références

[1] Rondeau, V., Mazroui, Y. and Gonzalez, J. (2012). frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, **47**(4), 1-28

[2] Rondeau, V. and Gonzalez, J. R. (2005). frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Comput Methods Programs Biomed*, **80**(2), 154-64.

[3] Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V. and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, **8**(4), 708-721.

[4] Rondeau, V., Pignon, J.-P. and Michiels, S. (2011). A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. *Statistical methods in medical research*

[5] Mazroui, Y., Mathoulin-Pellissier, S., MacGrogan, G., Brouste, V. and Rondeau, V. (2013). Multivariate frailty models for two types of recurrent events with an informative terminal event: Application to breast cancer data. To appear in *Biometrical journal*

The Dataset Project: Handling survey data in R

E. Rousseaux[a] and **G. Ritschard**[a]

[a] NCCR LIVES
Institute for Demographic and Life Course Studies
University of Geneva, Switzerland
emmanuel.rousseaux@unige.ch

Mots clefs : Survey, Data Management, Data processing, Data analysis, Panel Data.

Population studies strongly rely on survey data. To meet the needs of recent research questions in social sciences, data collected have becomed in the past decades more and more complex, such as longitudinal data, network data and spatial data. These high volumes of structured data complicate the task of both documenting data and manipulating data, as for example when we want to prepare data for a specific study. There is a need for specific tools to assist the user in handling these complex data. The Dataset software is an effort in this direction. It aims at providing a framework for handling survey data in R, especially network and biographical data. More precisely, the software aims at facilitating the management of survey data by providing researchers in social sciences with high-level tools for storing, documenting, sharing, exploring and recoding survey data in a secure and efficient way. This initiative, conducted within the NCCR LIVES project, targets mainly life course data and especially data types collected and used within the NCCR LIVES project. Thus, the current roadmap includes the development of the framework to support (1) cross-sectional data, (2) network data with a specific handling of demographic data from people cited in the network of each respondent, and (3) panel data organized in successive waves. The software comes in the form of an R package which is currently available on the R-Forge platform. R is a powerful statistical tool, freely available and multi-platform which is nowadays more and more often used in the social sciences as an alternative to classical commercial software (SPSS, SAS, Stata). As R is open-source, a lot of researchers in methods appreciate to be able to share their work through this software. As a consequence, most of the recent state-of-the-art methods are available in R and many of them in R only. This is especially the case for the newest tools for life course analysis (e.g. the `TraMineR` package for life course sequence analysis and the `ltm` package for latent class modeling). Moreover, working in R allows benefiting from the numerous statistical procedures already optimized in R and taking advantages of the R powerful graphical capability.
From a general point of view, the Dataset software follows three goals:

- *Providing an efficient framework for storing and documenting complex survey data.* As a key point, the software aims at storing data together with the design of the survey within which data were collected. Thus, the data and the user manual describing the data are merged together. Among the different features provided to describe data we can mention the possibility of assigning short and long labels to variables and variable values, to refer each variable to its question number within the survey, to declare user-defined types of missing values, and to account for cross-sectional and longitudinal weights. Many important metadata can also be stored such as the population concerned by the survey, the used sampling method, the organization releasing the data, the user license type, etc. As all information is stored within the data object, we provide a method for such objects for generating a summary of the whole data base. The summary gives for each variable

its long label, the percent of valid cases and basic descriptive statistics. This summary can be directly exported as a PDF file and serves as a basic user manual of the data base that proves particularly useful for detecting errors and for sharing data with others.

- *Saving the scientist's time spent on data processing in favor of time devoted to the research question.* Preparing data for a study is often a very burden task. The Dataset software is intended to help the analyst in this task, allowing him to focus more quickly on the analysis. As data bases are generally large, the package provides a search function allowing to explore the whole data base and retrieve relevant variables for the study. It provides efficient tools for recoding categorical and quantitative variables. The Dataset software also provides support for handling missing values and allows to easily turn a missing value into a valid case and vice-versa. Furthermore, the software provides, for some classical statistical methods, front-ends especially designed for scientists in social sciences. These front-ends facilitate the scientist daily work within the R environment. As a key point, the software performs systematic data consistency checks to ensure that data were not altered during data preprocessing operations. When filtering out cases and using weights when available, the software also processes automatic checks to prevent the loss of representativeness with respect to control variables defined by the user.

- *Facilitating reproducible research.* Demographic and sociologic questions are generally complex and require a lot of work to be understood. Reproducible research, meaning attaching sufficient information about the performed data analysis to allow anyone to retrieve the same results, is a helpful methodology when studying social dynamics. Having the possibility to rerun an experiment made by other researchers, or by oneself several months ago, gives the possibility to verify, better understand, and pursue an already done work. The Dataset software works in this direction by tracing operations made on data, so that the user can find back previously performed operations. Furthermore, for each statistical method provided by the package, results can be printed in a PDF file which also provides all settings used for calibrating the method. Outputs are displayed with a "ready-to-publish" formatting, allowing to quickly focusing on result interpretation.

In addition of these tools for cross-sectional data, the proposed software solution provides efficient methods for handling panel data organized in successive waves such as in the Swiss Household Panel. The user can directly extract whole trajectories from the panel data without having to bother with extracting the same variable independently from each yearly wave. The software automatically checks for each variable that it shares the same missing values and valid cases across years. By specifying ".." in place of the two year digits in the variable names, the user can extract a whole sequence in a single step. Likewise, the user can recode or merge values, or turn a missing value into a valid case directly for all waves where the variable exists. There also is a method for exporting a trajectory as a sequence object ready to be analyzed with the `TraMineR` package.

References

[1] Gabadinho, A., Ritschard, G., Müller, N. S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software, **40**(4), 1-37.
[2] Dimitris Rizopoulos (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, Journal of Statistical Software, **17**(5), 1-25.

Package SSDDA : Sample Size Determination and Data Analysis in the context of continuous co-primary endpoints in clinical trials.

P. Delorme[a], **P. Lafaye de Micheaux**[a], **B. Liquet**[b], and **J. Riou**[b,c]

[a] Département de mathématiques et de statistique
Université de Montréal
2920, chemin de la Tour, Montréal, Québec H3T 1J4, CANADA
lafaye@dms.umontreal.ca

[b]Equipe Biostatistique
ISPED
Université de Bordeaux - CR INSERM U897
146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE
benoit.liquet@isped.u-bordeaux2.fr

[c]Equipe Biometrie
Danone Research
RD 128, Avenue de la Vauve, 91767 Palaiseau Cedex, FRANCE
jeremie.riou@isped.u-bordeaux2.fr

Key-Words : Sample Size Determination, Co-primary endpoints, Multiple testing, Clinical trials

Nowadays, in clinical research, in order to capture a multi-factorial effect of some product, it is increasingly common to define multiple co-primary endpoints and then testing simultaneously a finite number of associated null hypotheses denoted $\mathcal{H}_0^k$, $k = 1, \ldots, m$. In this context where many hypotheses are tested, and each individual test has a specified Type I error probability, the probability that at least some Type I errors (false rejections or false positives) are committed increases with the number of hypotheses. Multiple hypothesis testing methods have been proposed for dealing with this problem. The most common one in clinical trials is undoubtedly the single step Bonferroni procedure. That could be explain by its ease of use and also for its control of the familywise error rate (FWER), which is defined as the probability of one or more false rejection among the family of the m hypotheses $\mathcal{H}_0^k$. This procedure is conservative (lead to wrongly "accepting" the null hypothesis) and might lead to biased test decisions, as information about correlations of the end points is not exploited.
In the context of "at least one win" continuous co-primary endpoints the aim of the work [1] is to provide sample size calculation methods, as well as corrections for Type-I errors probabilities based on a global method with a multivariate linear model or on an individual method involving a union-intersection procedure which controls the FWER and takes into account correlations among endpoints.
In the context of "at least r win" continuous co-primary endpoints, no procedures of sample size computation are developed. Therefore the aim of this work consists in providing a method which permits the sample size computation for single step (e.g. Bonferroni) and stepwise (e.g. Holm and Hochberg) procedures commonly used in clinical research. This choice will allow to seamlessly integrate this work within current clinical practice.
To facilitate the use of all these methods we have developed an R package SSDDA, which we

will present.

References

[1] Lafaye de Micheaux, P., Liquet, B., Marque, S., Riou, J. (2013). Power and sample size determination in clinical trials with multiple primary continuous correlated end points. To appear in *Journal of Biopharmaceutical Statistics*.

R as a sound system

J. Sueur[a]

[a]Muséum national d'Histoire naturelle
UMR CNRS-MNHN 7205 Origine Structure et Evolution de la Biodiversité
45 rue Buffon, 75005 Paris, France
sueur@mnhn.fr

Mots clefs : time series, acoustics, sound analysis, audio, time-frequency, sonification

Sound can be perceived everywhere at any time. In science, sound can be found in acoustics and in several other scientific and engineering disciplines, like musical acoustics, linguistics, speech and hearing sciences, psychoacoustics, bioacoustics, geology, noise control, and vibrations monitoring.

Handling sound with `R` is a rather easy task thanks to dedicated object classes directly deriving from `.wav` or `.mp3` audio file formats. It is possible to record, import, modify, and export audio objects with the packages `tuneR` [1] and `audio` [2]. Playing back a sound is mainly achieved by calling an external media player.

Sound can be analysed using `tuneR` and `seewave` [3] packages. These packages offer a set of complementary functions that can be used to extract and compare relevant amplitude, temporal, phase, and frequency parameters. For instance, Linear Predictive Coding, Fourier decomposition, Hilbert transform, cepstral analysis, or zero-crossing estimation are some of the techniques available to describe the frequency content of a sound. Figure 1 shows an example of a sound visualisation through a spectrographic representation with the tracking of both fundamental and dominant frequency bands. Correlation and distance functions can also be used to assess differences between pairs of amplitude envelopes or pairs of frequency spectral profiles. The use of the package `seewave` is now particularly important in bioacoustics – a discipline of life sciences focusing on animal sound – as it allows batch analyses of numerous audio files. This was in particular the case of studies that monitored the soundscapes of tropical forests and that returned up to 90,000 audio files to be analysed.

In addition to analysis, `tuneR` and `seewave` can generate sound by sinusoidal synthesis. Combined with arithmetic operations, frequency filtering and amplitude shaping, sinusoidal synthesis is simple but efficient and can produce signals that copy relatively well natural signals. Another side of sound synthesis is now available thanks to two new packages, `playitbyr` [4] and `audiolyzR` [5], that are dedicated to sonification [6]. Sonification is a type of auditory display aimed to communicate information. If visualization is the process of mapping data onto a graphic, then sonification can be seen as the process of transforming data into non-speech sound. In practical this means that you can listen to the crabs or iris data sets. `playitbyr`, which follows `ggplot2` syntax, maps data onto sonic parameters like pitch, tempo, and rhythm. Similarly, `audiolyzr` can generate audio representations of common plots accompanied with a pop-up panel to control interactively main sound parameters.

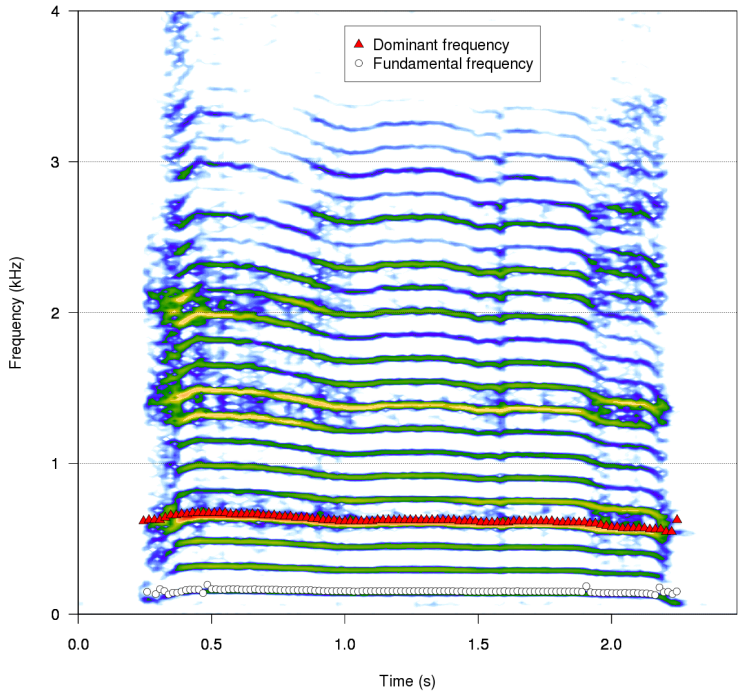


Figure 1: A call produced by a bird (*Vanellus vanellus*): a spectrographic representation with fundamental and dominant frequencies tracked. Produced with the `seewave` functions `spectro()`, `dfreq()`, and `fund()`. Electronic version with colours.

Références

[1] Ligges U. (2011). tuneR: analysis of music. R package version version 0.4-1. http://r-forge.r-project.org/projects/tuner/

[2] Urbanek S. (2011). audio: audio interface for R. R package version 0.1-4. http://CRAN.R-project.org/package=audio

[3] Sueur J., Aubin T., Simonis C. (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, **18**, 213-226.

[4] Brown E. (2012). playitbyr: representing and exploring data through sound. R package version 0.2-1. http://CRAN.R-project.org/package=playitbyr

[5] Stone E., Garrison J. (2013). audiolyzR: give your data a listen. R package version 0.4-9. http://CRAN.R-project.org/package=audiolyzR

[6] Hermann T., Hunt A., Neuhoff J. G. (2011). *The sonification handbook*. Logos Publishing House, Berlin, 586 p.

L'approche par comparaison de modèles avec R2STATS dans l'enseignement des statistiques en sciences humaines

Yvonnick Noël[a]

[a]Département de Psychologie
Université Européenne de Bretagne, Rennes 2
Place du Recteur Henri Le Moal, 35043 Rennes Cedex
yvonnick.noel@uhb.fr

Mots clefs : Interfaces graphiques pour R, GLM(M), facteur de Bayes, comparaison de modèles.

L'enseignement des statistiques aux étudiants en sciences humaines pose des problèmes spécifiques. Le public concerné est hétérogène et peu de nos étudiants montrent un goût affirmé pour la ligne de commande. Nous expérimentons depuis plusieurs années un enseignement basé sur la comparaison de modèles à l'aide d'interfaces graphiques pour `R`. La première, nommée `R2STATS` (Noël, 2012a) est une interface aux fonctions `glm()` de la librairie `base` et `glmer()` de la librairie `lme4`. Elle permet de construire facilement une séquence de GLM (ou de GLMM) emboîtés, d'en obtenir une représentation graphique automatique et de les comparer entre eux par des statistiques traditionnelles (F de Fisher, χ^2 sur la réduction de la déviance). La deuxième, nommée `AtelieR` (Noël, 2012b) inclut des outils de décision bayésiens plus récents (le facteur de Bayes) et des fonctions avancées de recherche automatique du meilleur modèle.

Ces interfaces encouragent l'approche qui consiste à traduire une question psychologique en une mise en concurrence de plusieurs modèles possibles. Dans cette approche, le point de départ est toujours le choix d'un modèle de distribution (conditionnel) sur la variable dépendante. C'est ce choix, argumenté théoriquement ou empiriquement (par un test d'ajustement), qui définit le type de problème statistique. Par exemple, la comparaison des résultats obtenus dans deux conditions expérimentales différentes ne se traduit par une comparaison de moyennes que dans le cadre d'une hypothèse de normalité et d'homogénéité des variances, car ce sont ces deux conditions qui réduisent la comparaison des modèles à une simple comparaison de paramètres de moyennes. Sous `R2STATS`, l'utilisateur définit d'abord son modèle de distribution (Figure 1), et les statistiques usuelles apparaissent dans le processus d'estimation des paramètres du modèle choisi (T de Student sur un coefficient), ou dans une comparaison de modèles concurrents (F de Fisher, χ^2 sur la réduction de la déviance).

L'interface `AtelieR` (Figure 2) implémente la même chose, dans une perspective bayésienne, en cherchant automatiquement le modèle le plus probablement vrai, sur un ensemble supposé exhaustif, pour les modèles binomiaux, multinomiaux et gaussiens (Noël, 2012b, 2013).

Références

[1] Noël, Y. (2013). *Psychologie statistique avec R*, coll. Pratique R, Paris : Springer.
[2] Noël, Y. (2012a). R2STATS: A GTK GUI for fitting and comparing GLM and GLMM in R. R package version 0.68-31. http://CRAN.R-project.org/package=R2STATS
[3] Noël, Y. (2012b). AtelieR: A GTK GUI for teaching basic concepts in Bayesian statistical inference. R package version 0.23. http://CRAN.R-project.org/package=AtelieR

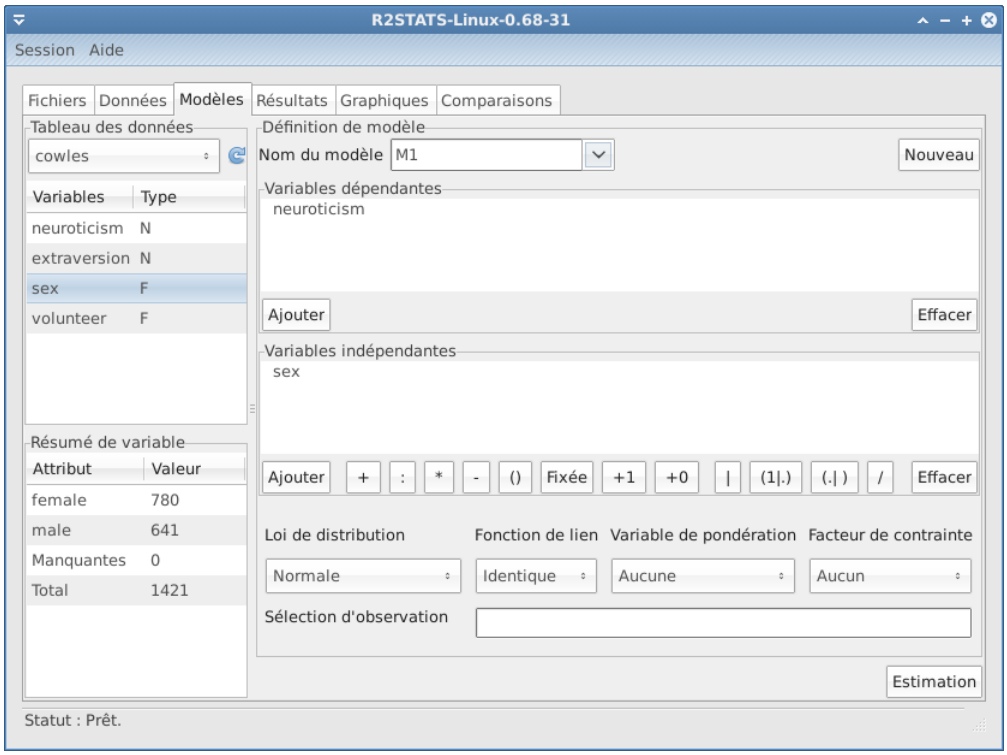


Figure 1: Définition de GLM sous R2STATS

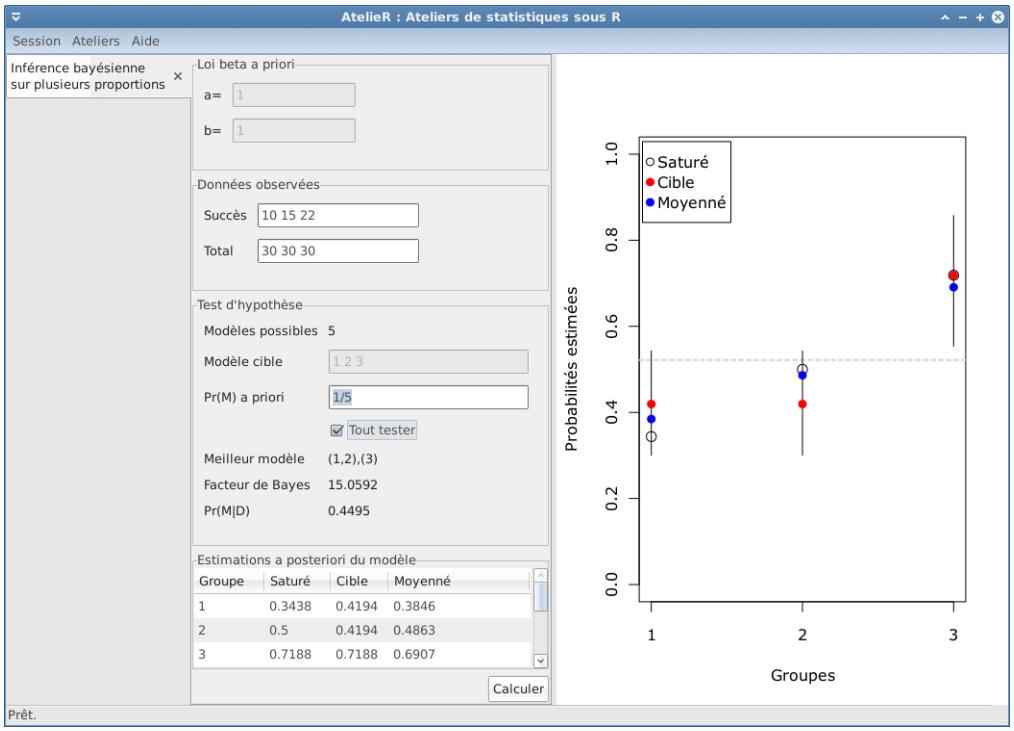


Figure 2: Comparaison bayésienne de proportions binomiales

Un site web d'enseignement R automatisé et à gestion partagée

S. Penel[a], S. Dray[a], A.B. Dufour[a], J. Lobry[b], et S. Mousset[a]

[a] Laboratoire de Biométrie et Biologie Evolutive, UMR 5558
CNRS, Université Claude Bernard, FST, Département de Biologie
43, Bd du 11 novembre 1918, 69622 Villeurbanne cedex
{simon.penel,stephane.dray,anne-beatrice.dufour,sylvain.mousset}@univ-lyon1.fr

[b] Institut National de la Police Scientifique
Ministère de l'Intérieur
31, av. Franklin Roosevelt, BP 30169 69134 Ecully cedex
jean.lobry@interieur.gouv.fr

Mots clefs : Statistique, Biologie, Enseignement.

Le site web «Enseignements de Statistique en Biologie» (http://pbil.univ-lyon1.fr/R) propose aux étudiants et aux utilisateurs de R plusieurs milliers de pages de supports pédagogiques : cours, TD, exercices, présentations, etc. L'aspect du site et son contenu sont intégralement gérés par les enseignants et le contenu R est vérifié automatiquement sur une base journalière.

Structure des documents
Les supports pédagogiques sont tous définis par 3 fichiers : un fichier en Sweave, un document PDF et un fichier de texte informatif. Le fichier en Sweave [1] contient le code LaTeX et le code R, le fichier PDF résulte de la compilation de ce fichier et le fichier de texte décrit la place du document dans la hiérarchie du site : type de document (cours, TD, exercices, etc.), thématique du document.

Vérification automatique journalière
Tous les codes Sweave de tous les documents du site sont vérifiés automatiquement chaque nuit grâce à un service démon (cron) en vue d'assurer la compatibilité entre les codes R proposés d'une part et les versions courantes de R et des packages utilisés d'autre part. Le package ade4 [2] bénéficie d'une attention particulière : les codes Sweave sont vérifiés pour la version de développement de ade4 (proposée sur la forge R) ainsi que pour la version courante de ade4 (proposée par le CRAN). Une page du site d'enseignement donne le résultat de la vérification journalière pour tous les codes R de toutes les fiches, ainsi qu'un accès au log de l'exécution et à l'historique des modifications du document.

Structure et maintenance du site web
La structure et la maintenance devaient répondre à plusieurs contraintes : partage des documents, conservation de l'historique des modifications, uniformisation de la mise en page et du format des documents, possibilité pour tous les enseignants de mettre en ligne les documents ou d'en supprimer l'accès, possibilité pour tous les enseignants d'ajouter de nouveaux documents, de modifier la structure du site et la hiérarchie des documents. L'aspect initial du site crée par Daniel Chessel a été conservé.

Nous avons choisi d'utiliser le logiciel de gestion de version « subversion (svn) » [3] pour traiter à la fois le contenu (les pages d'enseignements) et le contenant (la structure du site web).

La gestion du site est organisée de la manière suivante :

- un dépôt svn central,
- une copie svn locale pour chaque utilisateur : mise à jour à partir du dépôt central et envoi des modifications locales au dépôt central,
- une copie svn locale pour le web : c'est la version visible via la façade web. Ce dépôt peut seulement être mis à jour, ceci via une interface web. L'interface web permet aussi de mettre en ligne ou non les documents puis de reconstruire la façade web. C'est aussi cette copie locale qui est vérifiée automatiquement.

Exemple d'utilisation
L'utilisateur veut ajouter une nouvelle fiche TD dans une nouvelle section du site web :

1) l'utilisateur met à jour sa copie locale,
2) il modifie la structure du site web,
3) il ajoute un nouveau document (ajout de 3 fichiers : latex, pdf, info),
4) il soumet les modifications au dépôt central.

puis, sur l'interface web :

5) il met à jour la copie locale du site web,
6) il met en ligne le nouveau document,
7) il reconstruit la façade.

Références
[1] F. Leisch. « Sweave: Dynamic generation of statistical reports using literate data analysis. » In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*, pages 575-580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9
[2] S. Dray et A.B. Dufour **2007** « The ade4 package: implementing the duality diagram for ecologists » Journal of Statistical Software 22(4):1-20.
[3] http://subversion.apache.org/

Génération automatique de documents pédagogiques avec R pour l'enseignement et l'évaluation des étudiants.

S. Mousset

Laboratoire de Biométrie et Biologie Évolutive
CNRS UMR 5558 & Université Lyon 1
43, bd du 11 novembre 1918
F-69622 Villeurbanne – France
sylvain.mousset@univ-lyon1.fr

Mots clefs : Pédagogie, génération automatique de documents, questionnaires à choix multiples.

Dans le cadre de l'enseignement des mathématiques appliquées à la biologie, nous avons entrepris un effort de mise en commun et d'uniformisation des supports d'enseignement dans le but de générer simplement et efficacement deux types de documents pédagogiques :
• des fascicules d'exercices,
• des Questionnaires à Choix Multiples.

Dans ce cadre, nous utilisons la combinaison logicielle suivante :
• `R` [1] est utilisé en combinaison avec `Sweave` pour générer des sources LaTeX d'exercices et de questions à choix multiples dont les valeurs numériques peuvent être générées par `R`.
• `AMC` [2] utilise les sources LaTeX pour la production de questionnaires à choix multiples randomisés et leur correction automatique.
• LaTeX [3] est utilisé pour la composition des documents.
• `subversion` [4] permet de maintenir un dépôt versionné des documents utilisés.

Je présenterai l'architecture du dépôt que nous utilisons, ainsi que le processus de production des documents. À partir d'exemples d'exercices et de questions de QCM, je montrerai l'intérêt de l'utilisation de `R` combiné à `AMC` pour générer des questionnaires à énoncés pseudo-aléatoires corrigés automatiquement. Ces outils sont utilisés à l'Université Lyon 1 depuis l'automne 2009 dans une unité d'enseignement de première année de licence (400 étudiants par semestre) et ont été progressivement adoptés par d'autres UE depuis.

Références

[1] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
[2] Bienvenüe, A. (2008). Auto Multiple Choice. Conception et correction automatisée de QCM. URL http://home.gna.org/auto-qcm
[3] Lamport, L. (1986). LaTeX: A Document Preparation System. Addison-Wesley. ISBN 0-201-15790-X.
[4] Collins-Sussman, B., Fitzpatrick, B.W., Pilato, C.M. (2008). Version control with subversion. URL http://subversion.apache.org/

Pistes de réflexion pour la mise en œuvre d'un enseignement à distance sur le test du khi carré d'indépendance pour des étudiants en master de sciences de l'éducation

Mehdi Khaneboubi [a]

[a] Laboratoire EMA (EA 4507)
IUFM de l'université de Cergy-Pontoise
Site universitaire de Gennevilliers
Bureau c231,
ZAC des Barbanniers
Avenue Marcel Paul
92230 Gennevilliers
mehdi.khaneboubi@-cergy.fr

Mots clefs : initiation, enseignements à distance, sciences de l'éducation, khi carré.

Contexte

Cette contribution présente la mise en œuvre d'un enseignement d'initiation à l'analyse de données avec *R* pour des étudiants en sciences humaines et sociales dans un cours à distance. Ce cours de master professionnel, aussi dupliqué dans un master recherche de sciences de l'éducation, fait partie d'un diplôme conduit conjointement avec l'agence universitaire de la francophonie (AUF) et deux autres institutions d'enseignement supérieur en Belgique et en Suisse.

Les étudiants sont principalement localisés en Afrique et souvent, ne disposent pas de connexions à internet robustes ni d'accès à des bibliothèques universitaires. Tenant compte de ces contraintes, le tchat (clavardage) s'est imposé comme mode d'interaction pédagogique avec les étudiants, pour compléter des supports pédagogiques en ligne. La durée des enseignements est d'environ 6 semaines à raison d'une à deux séances de tchat par semaine. Dans ces diplômes, cet enseignement est le seul à traiter de méthodes quantitatives. C'est aussi le seul enseignement dans lequel les étudiants auront l'occasion de s'initier à la programmation à l'exception d'un autre cours de master professionnel. Les étudiants sont en majorité des enseignants, des formateurs ou des prescripteurs intermédiaires se destinant soit à évoluer dans leurs institutions soit à un doctorat.

Dans ce contexte, quels atouts présente *R* par rapport à d'autres logiciels de traitement statistiques ? Quelle progression didactique est pertinente ? Comment articuler instrumentation logicielle et manipulation des savoirs statistiques ? Quelles difficultés peut-on identifier du point de vue des apprenants ? Au travers d'une analyse des tchats, nous verrons quels sont les principaux obstacles et les détails clés à considérer pour mettre œuvre cet enseignement.

Mise en œuvre

Dans notre contexte *R* à l'avantage de permettre de communiquer les scripts par tchat et de les examiner collectivement avec les étudiants. De plus, les ressources francophones disponibles en ligne sont doublement adaptées à l'isolement d'apprenants en formation à distance et à des étudiants n'ayant qu'un accès difficile à des bibliothèques universitaires.

D'un point de vue didactique, un certain nombre de préalables doivent être appréhendés pour que les étudiants puissent interpréter les informations produites par le script suivant :

```
# importation des données, dans les deux premières colonnes du fichier
sont des variables qualitatives
read.csv2("variables.csv" , sep=";" , header=TRUE , na.strings="NA")->bdd

# résumé de la base
summary(bdd)
head(bdd)
tail(bdd)

# calcul de l'indicateur de khi deux et de la p-value pour le tableau
constitué par les colonnes 1 et 2 de l'objet « bdd »
chisq.test(table(bdd[,1] , bdd[,2]) , correct = FALSE)->khideux

# tableau des effectifs observés
khideux$observed

# tableau des effectifs théoriques
khideux$expected

# racine carré du tableau de khi deux
khideux$residuals
```

D'une part, il faut les familiariser progressivement avec des manipulations techniques périphériques : installation du logiciel, gestion des formats de fichiers, repérage du répertoire de travail…
Ensuite avec le fonctionnement de *R* : comprendre ce qu'est un objet, distinguer un objet et une fonction, *R* ne répond pas lorsque tout fonctionne, *R* répond des choses incompréhensibles…
Et enfin avec les particularités du test de khi-deux d'indépendance : hypothèse d'indépendance, loi de khi deux, p-value, interprétation des résultats etc.

Références

[1] Barnier, J. (2008). *R pour les sociologues (et assimilés)*. Consulté à l'adresse http://alea.fr.eu.org/j/intro_R.html

[2] Cibois, P. (2007). *Les méthodes d'analyse d'enquêtes*. Paris: Presses Universitaires de France.

[3] Haspekian, M. (2005). An « Instrumental Approach » to Study the Integration of a Computer Tool Into Mathematics Teaching: the Case of Spreadsheets. *International Journal of Computers for Mathematical Learning*, *10*(2), 109-141.

[4] Roditi, E. (2009). Un tableur grapheur pour enseigner les statistiques en sciences humaines et sociales. In G.-L. Baron, É. Bruillard, & L.-O. Pochon (Éd.), *Informatique et progiciels en éducation et en formation* (p. 257‑275). Consulté à l'adresse http://halshs.archives-ouvertes.fr/halshs-00609639

[5] Tort, F., Blondel, F. M., & Bruillard, É. (2008). Spreadsheet knowledge and skills of French secondary school students. *Informatics Education-Supporting Computational Thinking*, 305–316. Consulté à l'adresse http://www.springerlink.com/index/U5457121103G7645.pdf

De la biologie à l'algèbre linéaire ... en passant par R
Expérimenter la notion de projection

A.B. Dufour[a], **S. Dray**[a], **J.R. Lobry**[b] and **J. Thioulouse**[a]

[a]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558
CNRS, Université Claude Bernard, FST, Département de Biologie
43, Bd du 11 novembre 1918, 69622 Villeurbanne cedex
anne-beatrice.dufour@univ-lyon1.fr

[b]Institut National de la Police Scientifique
Ministère de l'Intérieur
31, av. Franklin Roosevelt, BP 30169 69134 Ecully cedex
jean.lobry@interieur.gouv.fr

Mots clefs : Enseignement, Biologie Humaine, Algèbre linéaire, Projection.

L'enseignement de la statistique auprès des biologistes se confond, à Lyon, avec l'histoire du laboratoire de Biométrie et de son fondateur J.M. Legay. Dès le début des années soixante, ce dernier réunit biologistes et mathématiciens pour initier un dialogue et développer de nouvelles méthodes [1]. Cette impulsion conduit à la création d'enseignements intégrés liant biologie, statistique et informatique. L'objectif de cette communication est de montrer l'apport du logiciel R dans cette relation triangulaire [2].

Une partie de la statistique comme l'analyse des données relève de l'algèbre linéaire et plus particulièrement de la notion de projection. Celle-ci prend des formes différentes selon la problématique posée : (i) les variables étudiées jouent des rôles asymétriques (*i.e.* explicatives ou à expliquer) comme en régression linéaire simple, (2) les variables étudiées jouent un rôle identique comme dans l'analyse en composantes principales. Cette projection est l'essence même de la méthode mais son formalisme mathématique peut la rendre difficile à comprendre.

C'est pourquoi la méthode et la notion de projection sont exposées à partir de la donnée brute. Elles se visualisent, s'expérimentent, se révèlent. L'exemple proposé porte sur les mesures de la stature (taille, en cm), de l'empan de la main dominante (empan1, en cm) et de l'empan de la main non dominante (empan2, en cm) réalisées sur 168 étudiants (`data(survey)` de la librairie `MASS`). L'empan est la distance entre l'auriculaire et le pouce, le poignet et la main étant posés, à plat sur une table, les doigts écartés au maximum.

Il existe une relation linéaire entre l'empan et la taille d'un individu. Cette relation peut être modélisée par une droite dite droite de régression. L'objectif est de trouver les valeurs de l'ordonnée à l'origine et de la pente qui minimisent la somme des carrés des écarts entre la valeur de l'empan et son estimation par le modèle (sa projection parallèlement à l'empan, variable à expliquer). Avec le logiciel R, une fonction des deux paramètres (pente et ordonnée à l'origine) peut être construite et l'étudiant peut essayer de rechercher la droite optimum (Figure 1, [3]).

Mais il est rare de n'avoir que deux mesures à mettre en relation. Un des objectifs de la morphométrie est de séparer la taille globale d'un individu de sa forme. L'idée est alors de prendre l'ensemble des mesures, de ne privilégier aucune variable et de rechercher cette taille globale. La solution est le premier axe d'une analyse en composantes principales c'est-à-dire de la droite qui maximise la variance projetée ou qui minimise la somme des carrés des écarts entre un individu et sa projection orthogonale. Avec le logiciel R et la fonction `plot3d`, l'étudiant peut faire tourner le nuage de points (liant taille et empans) jusqu'à faire apparaître une direction la plus étendue possible (Figure 2, [4]).

Enseigner la statistique par le formalisme mathématique éloigne de la donnée. Pour que l'étudiant comprenne et s'approprie une méthode, l'enseignant se doit d'être pragmatique. L'échange s'opère autour de la visualisation de la donnée et de la méthode. Les outils proposés par le logiciel R permettent d'initier ce dialogue.

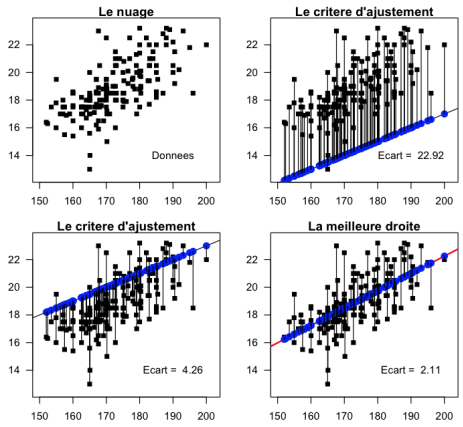


Figure 1 : Différentes expressions de la relation liant la taille et l'empan de la main dominante

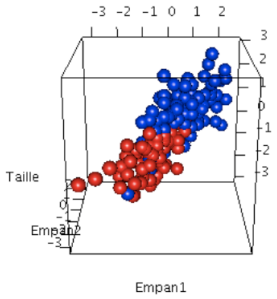


Figure 2 : Une représentation du nuage des 168 étudiants en dimension 3

Références

[1] Legay, J.M. (2004) L'interdisciplinarité vue et pratiquée par les chercheurs en sciences de la vie. *Nature, Sciences et Sociétés*, **12**, 63-74

[2] Dufour, A.B. (2012) La part du logiciel R dans l'enseignement de la statistique en biologie. Le site Web de Lyon. *Statistique et Enseignement*, **2**(2), 41-47

[3] Dufour, A.B., Lobry, J.R., Chessel, D., Rochette, N. (2012) De la stature chez l'Homme ... à la taille des cerveaux chez les mammifères. Réversion, Régression, Corrélation.
`http://pbil.univ-lyon1.fr/R_svn/pdf/bem3.pdf`

[4] Dufour, A.B., Lobry, J.R. (2011) Initiation à l'analyse en composantes principales.
`http://pbil.univ-lyon1.fr/R_svn/pdf/tdr601.pdf`

metaRNASeq : un package pour la méta-analyse de données RNA-seq

G. Marot[a,b], **F. Jaffrézic**[c] and **A. Rau**[c]

[a]EA2694 Centre d'Etudes et de Recherche en Informatique Médicale
Université Lille 2
1 place de Verdun, 59045 Lille cedex
guillemette.marot@univ-lille2.fr

[b]Equipe Projet Inria MODAL
Inria
40 avenue Halley - Bat A , 59655 Villeneuve d'Ascq cedex
guillemette.marot@inria.fr

[c]UMR1313 Génétique Animale et Biologie Intégrative
INRA
Domaine de Vilvert - 78350 Jouy-en-Josas cedex
florence.jaffrezic@jouy.inra.fr

Mots clefs : Biostatistique, méta-analyse, analyse différentielle, RNA-seq, transcriptomique, séquençage haut débit

Les techniques de séquençage à haut débit telles que le RNAseq sont de plus en plus utilisées pour les analyses de données transcriptomiques. Cependant, en raison du coût encore élevé des expériences, peu de réplicats biologiques sont inclus dans les études, ce qui affecte la capacité de détection des vrais transcrits différentiellement exprimés. Il est probable qu'avec la diminution du coût du séquençage, des expériences soient reconduites pour répondre à des questions déjà posées et gagner en sensibilité en rajoutant des réplicats. Il est donc nécessaire de développer des techniques qui puissent analyser conjointement les résultats d'analyse différentielle de différentes études. Ces méthodes doivent tenir compte de la variabilité biologique et technique à l'intérieur de chaque expérience ainsi que de l'effet inter-études [1]. Nous avons comparé les méthodes de combinaison de p-values déjà utilisées dans des analyses de puces à ADN à un modèle linéaire généralisé (GLM) incluant un effet étude. Ces comparaisons à la fois sur des jeux de données réels et des simulations ont confirmé que le GLM avec effet étude se comportait très bien quand peu d'études étaient disponibles et que l'effet étude était faible. Elles ont aussi montré que les techniques de méta-analyse étaient plus performantes que le GLM étudié quand la variabilité entre études était grande et le nombre d'études important.

Le package metaRNAseq, disponible sur R-Forge, implémente les techniques de méta-analyse présentées dans [1]. Après avoir redonné les principaux résultats du papier correspondant, nous présenterons rapidement la vignette de ce package en insistant sur les différences entre les techniques précédemment utilisées pour les puces à ADN [2] et celles développées pour le séquençage [1]. Ces différences concernent notamment la gestion des conflits entre les gènes sous-exprimés dans une étude et sur-exprimés dans une autre. Nous porterons aussi une attention particulière sur les vérifications préliminaires à effectuer dans une méta-analyse de données RNA-seq pour se placer dans un cadre où les techniques implémentées sont effectivement les meilleures. En particulier, nous insisterons sur la nécessité d'observer des distributions de p-values uniformes sous l'hypothèse nulle dans chaque étude, ce qui est possible en utilisant la

méthode développée par [3].

Références

[1] Rau, A., Marot, G., Jaffrézic, F. (2013). Differential meta-analysis of RNA-seq data from multiple studies. In preparation.

[2] Marot, G. , Foulley, J.-L., Mayer, C.-D., Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses, *Bioinformatics*, **25**(20), 2692–2699

[3] Rau, A., Gallopin, M., Celeux, G., Jaffrézic, F. (2013). Independent data-based filtering for replicated high-throughput transcriptome sequencing experiments. Submitted.

jointSeg : Segmentation de données génomiques en cancérologie

M. Pierre-Jean[a,b] et **P. Neuvial**[b]

[a]EA 2694 Université Lille 2
Centre d'Etudes et de Recherche en Informatique Médicale
1 place de Verdun, 59045 Lille cedex
morgane.pierrejean@genopole.cnrs.fr

[b]UMR 8071 CNRS - Université d'Evry- INRA
Laboratoire Statistique et Génome
23 boulevard de France, 91037 Evry cedex
pierre.neuvial@genopole.cnrs.fr

Mots clefs : Bioinformatique, CNV, Cancer, Nombre de copies, Fraction d'allèle B, biostatistique, détection de ruptures.

L'identification des régions du génome où le nombre de copies d'ADN a été altéré dans les cellules cancéreuses permet de mieux comprendre la progression des tumeurs et de mettre en place des thérapies personnalisées. Nous nous sommes intéressés à la détection de ruptures dans les profils génomiques issus d'échantillons de cellules cancéreuses.

Le package `jointSeg` est disponible depuis janvier 2013 sur R-forge [1]. Il permet notamment :

1. de générer simplement des profils synthétiques réalistes, à l'aide d'un petit nombre de paramètres dont l'interprétation biologique est claire : la proportion de cellules tumorales, la longueur du signal, le nombre de ruptures ;

```
> library(jointSeg)
> data <- loadCnRegionData(platform="Affymetrix", tumorFraction=.5)
> set.seed(1)  ## for full reproducibility
> sim <- getCopyNumberDataByResampling(2e4, nBkp=4, regData=data)
```

2. l'utilisation de plusieurs méthodes de segmentation existantes via une interface unifiée :
   – approches exactes par programmation dynamique (`cghseg` [1]) ;
   – segmentation binaire (`CBS` [4], `PSCBS` [6])
   – régression pénalisée de type fused Lasso (`GFLARS` [2], portage en R d'un code Matlab)
   – modèle de Markov caché (`PSCN` [3]).
   Nous avons également implémenté une méthode que nous avons appelée `RBS` pour Recursive Binary Segmentation, et qui combine `CART` et la programmation dynamique [5] :

```
> resRBS <- PSSeg(data=sim$profile, K=20, flavor="RBS", profile=TRUE)
```

3. la représentation graphique des résultats (Figure 1) ;

```
> plotSeg(sim$profile, list(true=sim$bkp, est=resRBS$bestBkp))
```

4. l'évaluation des performances des différentes méthodes en fonction de la taille d'une zone de tolérance autour des vraies ruptures (non illustrée dans ce résumé pour des raisons de place).

1. `http://r-forge.r-project.org/R/?group_id=1562`

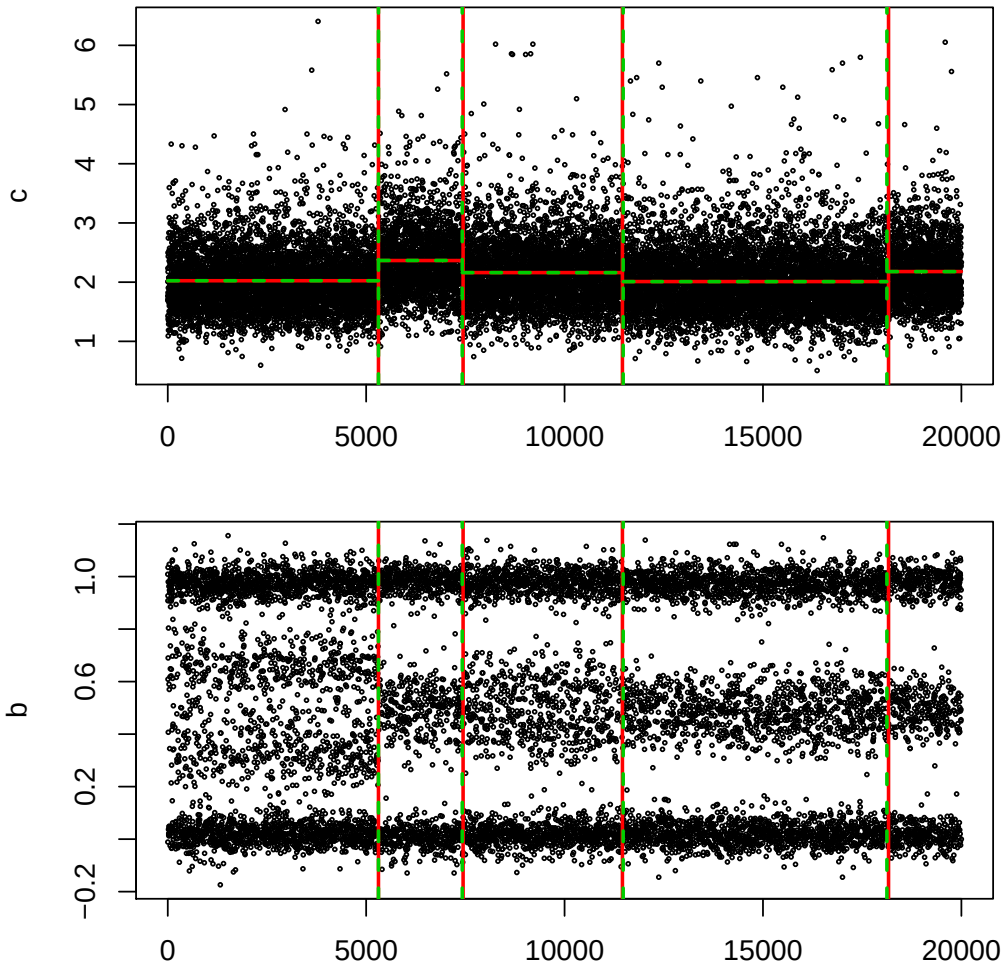


FIGURE 1 – Exemple de données synthétiques produites par le package `jointSeg`. Lignes verticales rouges : vraies ruptures ; lignes verticales vertes : points de ruptures identifiés par `RBS`.

Références

[1] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. Technical report, http ://arXiv.org/abs/1004.0887, 2010.

[2] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neutral Information Processing Systems*, 2010.

[3] Chen, H., Xing, H. and Zhang, N.R. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput Biol*,2011.

[4] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, (2004).

[5] Gey, S. and Lebarbier, E., Using CART to Detect Multiple Change Points in the Mean for Large Sample, *Statistics for Systems Biology research group*, (2008)

[6] Olshen, Adam B and Bengtsson, Henrik and Neuvial, Pierre and Spellman, Paul T and Olshen, Richard A and Seshan, Venkatraman E, Parent-specific copy number in paired tumor-normal studies using circular binary segmentation *Boinformatics*, (2011)

HTSFilter: An independent data-based filter for replicated high-throughput transcriptome sequencing experiments

A. Rau[a], **M. Gallopin**[a], **G. Celeux**[b] and **F. Jaffrézic**[a]

[a]UMR 1313 GABI
INRA
Jouy-en-Josas, France 78352
{andrea.rau, melina.gallopin, florence.jaffrezic}@jouy.inra.fr

[b]Inria Saclay – Île-de-France
Orsay, France 91405
gilles.celeux@math.u-psud.fr

Mots clefs : Independent filter, gene expression, RNA-seq, differential analysis.

Over the past five years, next-generation high-throughput sequencing (HTS) technology has become an essential tool for genomic and transcriptomic studies. In particular, the use of HTS technology to directly sequence the transcriptome, known as RNA sequencing (RNA-seq), has revolutionized the study of gene expression by opening the door to a wide range of novel applications. Unlike microarray data, RNA-seq data represent highly heterogeneous counts for genomic regions of interest (typically genes), and often exhibit zero-inflation and a large amount of overdispersion among biological replicates. As such, a great deal of methodological research has recently focused on appropriate normalization and analysis techniques that are adapted to the characteristics of RNA-seq data, particularly for the study of differential expression among experimental conditions.

Because a large number of hypothesis tests (typically in the thousands or tens of thousands) are performed for gene-by-gene differential analyses, stringent false discovery rate control is required at the expense of the power of an experiment to detect truly differentially expressed (DE) genes. To reduce this impact, data filters are often used in order to identify and remove genes which appear to generate an uninformative signal and have little chance of showing significant evidence of differential expression; only hypotheses corresponding to genes that pass the filter are subsequently tested, which in turn tempers the correction needed to adjust for multiple testing. However, in practice an arbitrary filtering threshold is typically fixed for RNA-seq data without accounting for the overall sequencing depth or variability of a given experiment, and little attention is paid to its impact on the downstream analysis.

In this work, we propose a Bioconductor package, `HTSFilter`, that implements a data-driven method to identify an appropriate filtering threshold for replicated RNA-seq data [1]. The main idea underlying this method is to identify the threshold that maximizes the filtering similarity among biological replicates, that is, one where most genes tend to either have normalized counts less than or equal to the cutoff in all samples (i.e., filtered genes) or greater than the cutoff in all samples (i.e., non-filtered genes). More precisely, we denote the observed read counts for all genes in sample j as $\mathbf{y}_j = (y_{gj})$, where $\mathcal{C}(j)$ is the experimental condition of sample j. After binarizing the data for a fixed filtering threshold s (1 if $y_{gj} > s$ and 0 otherwise), the Jaccard

		Sample j	
		Normalized counts $> s$	Normalized counts $\leq s$
Sample j'	Normalized counts $> s$	a	b
	Normalized counts $\leq s$	c	d

Table 1: Constants used to calculate the Jaccard index defined in Equation (1).

similarity between two biological replicates may be defined as follows:

$$J_s(\mathbf{y}_j, \mathbf{y}_{j'}) = \frac{a}{a+b+c}, \tag{1}$$

where a, b, and c are defined in Table 1. Because multiple replicates and conditions are typically available in HTS experiments, we extend the definition of the pairwise Jaccard index in (1) to a global Jaccard index by averaging the indices calculated over all pairs in each condition:

$$J_s^\star(\mathbf{y}) = \text{mean}\left\{J_s(\mathbf{y}_j, \mathbf{y}_{j'}) : j < j' \text{ and } \mathcal{C}(j) = \mathcal{C}(j')\right\}. \tag{2}$$

Finally, we identify the threshold $s^\star$ that yields the largest possible global Jaccard index:

$$s^\star = \underset{s}{\text{argmax}}\, J_s^\star(\mathbf{y}).$$

In practice, in the `HTSFilter` package we calculate the value of the global Jaccard index in (2) for a fixed set of threshold values and fit a loess curve through the set of points; the value of $s^\star$ is subsequently set to be the maximum of these fitted values.

In comparisons with alternative data filters regularly used in practice, we have demonstrated the effectiveness of our proposed method to correctly filter weakly expressed genes, leading to increased detection power for moderately to highly expressed genes. Interestingly, this data-driven threshold varies among experiments, highlighting the interest of the method proposed here. The `HTSFilter` package is compatible with a variety of data classes and analysis pipelines that have been proposed for RNA-seq data, including `matrix` and `data.frame` objects, the S4 class `CountDataSet` in the `DESeq` pipeline [2], and the S3 class `DGEList` in the `edgeR` pipeline [3]. A package vignette distributed with the `HTSFilter` package describes the use of the filtering method within each of these pipelines.

Références

[1] Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013) Independent data-based filtering for replicated high-throughput transcriptome sequencing experiments (submitted).
[2] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, 11(R106):1-28.
[3] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139-140.

Aspects de cartographie thématique pour les sciences sociales avec R

Joël Gombin[a]

[a]CURAPP – Université de Picardie Jules Verne
Faculté de Droit et de Science Politique
Pôle universitaire Cathédrale
10, placette Lafleur
BP 2716
80027 AMIENS Cedex
joel.gombin@u-picardie.fr

Mots clefs : Cartographie, Sciences sociales.

R n'est pas initialement un logiciel dédié à la cartographie, et alors que se développent d'ambitieux SIG en open-source (QGIS ou GRASS parmi les plus connus, ou plus récemment et plus orienté cartographie et moins SIG TileMill), l'idée de faire de la cartographie avec un logiciel de statistiques peut paraître incongrue. Toutefois, la souplesse de R en fait un outil de choix, en particulier en matière de géostatistiques (Bivand et al., 2008), et à la suie du développement du package `sp`, de nombreuses fonctionnalités cartographiques et d'analyse spatiale ont été développées. La popularisation de bases de données cartographiques telles que Google Maps ou OpenStreetMap ont également provoqué l'apparition de packages permettant l'accès à ces bases.

Toutefois, la prolifération des packages et fonctions permettant la production de cartes géographiques dans R n'a pas permis, jusqu'à présent, d'imposer un cadre de référence. Cela participe sans aucun doute de la richesse et de la souplesse offertes par R, mais dans le même temps rend plus complexe la réalisation de cartes pour le néophyte. La tentation peut alors être grande, y compris pour des praticiens de R, d'avoir recours à des logiciels spécialisés en cartographie et ainsi de dissocier l'analyse statistique de l'analyse cartographique – ce qui est clairement une pratique inefficace, souvent non-reproductible et qui peut être à l'origine d'erreurs.

Le *lightning talk* que je propose n'entend pas proposer un cadre unifié pour la cartographie, mais dresser un rapide état des lieux de l'existant pour les usagers en sciences sociales, et présenter des fonctions inédites de cartographie écrites en programmation orientée object (POO) qui permettent une approche modulaire et souple de la cartographie thématique. Cette approche devrait permettre la réalisation relativement aisée de cartes sophistiquées et soignées pour les chercheurs, enseignants et étudiants en sciences sociales. Les fonctions présentées seront disponibles sur `http://www.github.com/joelgombin`.

Références

Bivand Roger S., Pebesma Edzer J. et Rubio Virgilio Gómez, *Applied spatial data : analysis with R*, Springer, 2008, 379 p.

Nouvelles fonctionnalités du package `fitdistrplus`

ML. Delignette-Muller[a] and **C. Dutang**[b]

[a]Université de Lyon
Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive
VetAgro Sup, Campus vétérinaire de Lyon
1 avenue Bourgelat, 69280 Marcy l'Etoile
marielaure.delignettemuller@vetagro-sup.fr

[b]Université de Strasbourg
CNRS, UMR7501, Institut de Recherche Mathématiques Avancée
7 rue René Descartes, 67084 Strasbourg Cedex
dutang@math.unistra.fr

Mots clefs : ajustement de distributions, bootstrap, données censurées.

`fitdistrplus` est un package **R** dédié à l'ajustement de distributions paramétriques à des données univariées. Il propose diverses fonctions visant à faciliter le processus global de description d'une distribution empirique par une distribution paramétrique, incluant :

1. le choix de distributions candidates pour décrire les données,
2. l'ajustement de chacune des distributions candidates aux données,
3. la comparaison des ajustements en vue de choisir la distribution la plus adaptée,
4. le calcul, par bootstrap, de l'incertitude sur les paramètres estimés de la distribution choisie.

Plusieurs méthodes d'estimation des paramètres sont proposées dans le package : les plus courantes que sont la méthode du maximum de vraisemblance et la méthode des moments [1], mais aussi la méthode des quantiles [2] et la méthode de minimisation d'une statistique d'ajustement [3].

Une spécificité importante du package est la prise en compte de données de types variés. En effet, les fonctions relatives à l'ajustement de distributions par maximum de vraisemblance ont été adaptées pour permettre la prise en compte d'une part des données discrètes, et d'autre part des données censurées [4], quel que soit le type de censures (à droite, à gauche ou par intervalle).

Ce package a tout d'abord été développé pour une utilisation dans le cadre de l'appréciation quantitative des risques [5], mais les outils qu'il propose sont de nature à aider tout scientifique dans l'ajustement de distributions à des données observées. Depuis sa publication sur le CRAN en 2009, il a d'ailleurs été utilisé dans des domaines d'application très variés : appréciation quantitative des risques, épidémiologie, biologie moléculaire, génomique, bioinformatique, mathématiques financières et actuarielles, ...

Les divers retours des utilisateurs nous ont conduit récemment à développer de nouvelles fonctionnalités que nous souhaitons présenter. Sont en particulier maintenant disponibles de nou-

velles fonctions facilitant la comparaison des ajustements de plusieurs distributions sur un même jeu de données, tant au niveau graphique [1] (fonctions `denscomp`, `cdfcomp`, `ppcomp` et `qqcomp` respectivement pour les diagrammes en densité, en fréquences cumulées et les P-P plot et Q-Q plot) qu'au niveau numérique [6] (amélioration de la fonction `gofstat` pour le calcul des statistiques d'ajustement et des critères d'information). La fonction générique `quantile` a aussi été créée pour diverses classes **S3** d'objets définies dans le package : elle permet le calcul de quantiles à partir d'une distribution ajustée sur des données censurées ou non, ainsi que le calcul par bootstrap de l'incertitude sur ces quantiles estimés.

Références

[1] Cullen, A., Frey, H. (1999). *Probabilistic techniques in exposure assessment.* First edition. Plenum Publishing Co.

[2] Tse, Y. (2009). *Nonlife Actuarial Models: Theory, Methods and Evaluation.* International Series on Actuarial Science. Cambridge University Press.

[3] Luceno, A. (2006). Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics and data analysis*, **51**(2), 904-917.

[4] Helsel, D. (2005). *Nondetects and data analysis: statistics for censored environmental data.* First edition. Wiley.

[5] Pouillot, R.,Delignette-Muller, M.L. (2010). Evaluating variability and uncertainty separately in microbial quantitative risk assessment using two R packages. *International Journal of Food Microbiology*, **142**[3], 330-340.

[6] D'Agostino, R., Stephens, M. (1986). *Googness-of-fit techniques.* First edition. Dekker.

SesIndexCreatoR: Un package R pour la création et la visualisation d'indices socioéconomiques

B. Lalloué[a,b,c], **S. Deguen**[a,b], **J.-M. Monnez**[c], **C. Padilla**[a,b], **W. Kihal**[a], **D. Zmirou-Navier**[a,b,d] and **N. Le Meur**[a,e]

[a]EHESP Rennes, Sorbonne Paris Cité
Rennes, France
[b]Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085
Rennes, France
[c]IECL, Institut Elie Cartan Nancy, CNRS : UMR 7502
Université de Lorraine, INRIA BIGS, France
[d]Faculté de Médecine
Université de Lorraine, France
[e]UMR936 INSERM, Université de Rennes 1
Rennes, France

benoit.lalloue@ehesp.fr ; severine.deguen@ehesp.fr ; jean-marie.monnez@univ-lorraine.fr ; cindy.padilla@ehesp.fr ; wahida.kihal@ehesp.fr ; denis.zmirou@ehesp.fr ; nolwenn.lemeur-rouillard@ehesp.fr

Mots clefs : Analyse en composantes principales, Classification Ascendante Hiérarchique, Statut socio-économique.

Afin de prendre en compte les inégalités sociales de santé, il est fréquent d'utiliser des indices socio-économique qui synthétisent différents aspects du statut socio-économique (SES) à l'échelle de l'individu, du voisinage ou de la région. De nombreux indices existent déjà mais nombre d'entre eux utilisent un faible nombre de variables, des méthodes d'agrégation simples et/ou sélectionnent les variables uniquement depuis la littérature. Nous avons développé une nouvelle procédure de création d'indices SES [1] visant à sélectionner à partir d'un grand nombre de variables celles qui composeront l'indice et à les agréger à l'aide de techniques d'analyses de données. Cette procédure est plus complexe à mettre en oeuvre pour des non statisticiens que d'autres approches existantes, c'est pourquoi nous l'avons implémentée dans un package R nommé SesIndexCreatoR. Ce package vise à donner des outils aussi simples d'emploi que possible pour effectuer la procédure de création et exploiter ses résultats, tout en conservant les différentes possibilités et la flexibilité qu'elle offre.

Le package SesIndexCreatoR dépend des packages FactoMineR et class. En particulier, la plupart des fonctions d'analyse de données et de visualisation sont issues et adaptées du package FactoMineR. Les sources du package SesIndexCreatoR sont disponibles librement sur http://www.equitarea.org/documents/packages_1.0-0/. Le package est composé de trois fonctions principales, de leurs fonctions de visualisation et d'impression associées, et de plusieurs autres fonctions internes :

- La fonction SesIndex crée un indice socio-économique utilisant la procédure décrite ailleurs [1]. Il est possible de choisir l'ensemble de variables de départ, les potentielles variables "redondantes" (représentant une même notion socio-économique), les potentielles unités illustratives, la méthode de sélection (ACP ou AFM) ou l'étape de la procédure à réaliser.

Les résultats incluent l'indice final et tous les résultats des étapes intermédiaires.

- La fonction SesClassif crée des catégories socio-économiques basées sur un indice créé par la fonction SesIndex, en offrant la possibilité d'utiliser différentes techniques : CAH (avec consolidation par les k plus proches voisins ou non), quantiles ou intervalles de même taille. Les résultats incluent à la fois le jeu de données d'origine auquel a été ajouté la catégorie de chaque individu, ainsi que les résultats de la technique de classification.

- La fonction SesReport crée un fichier HTML contenant un rapport synthétisant les résultats des différentes étapes de création d'un indice effectuée avec la fonction SesIndex et, s'il y en a une, de la classification de l'indice faite avec la fonction SesClassif. La fonction SesReport permet aussi l'export d'un fichier CSV contenant les données d'origine, l'indice calculé et l'éventuelle classification.

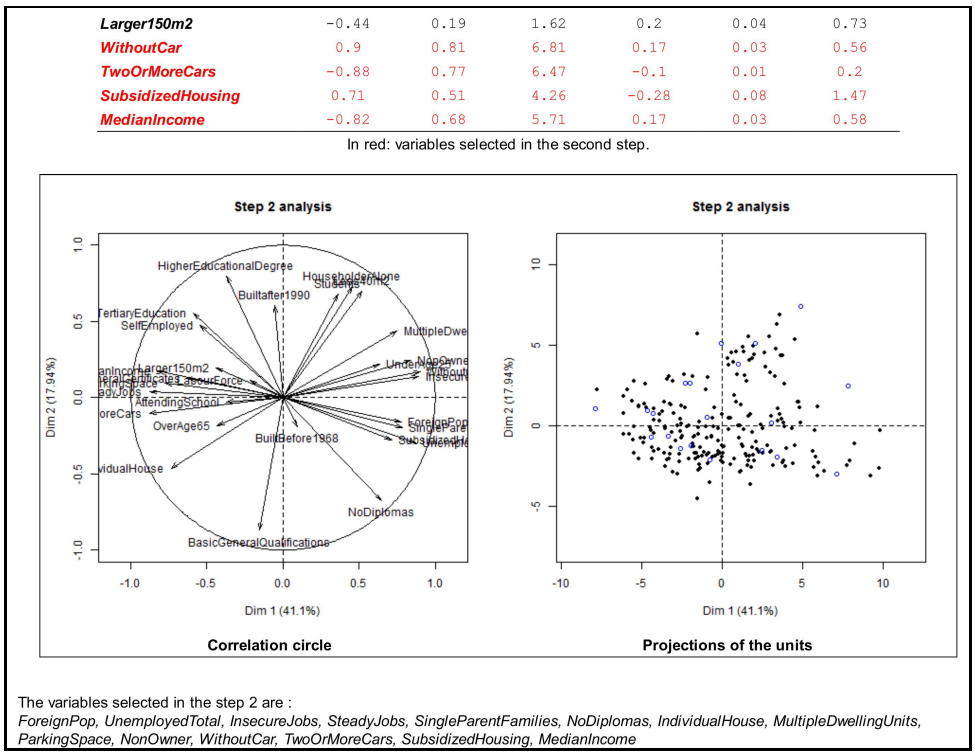


Figure 1: Extrait d'un rapport d'exemple généré par la fonction SesReport

Le package SesIndexCreatoR propose ainsi sous une forme "tout en un" l'ensemble des outils nécessaires pour appliquer, interpréter et utiliser la procédure de création d'un indice socio-économique développée, y compris pour des utilisateurs non statisticiens. Nous projetons par ailleurs d'étendre le package dans le futur en ajoutant entre autres de nouvelles méthodes de classification, davantage d'outils d'interprétation ou encore d'autres moyens de visualisation (comme de la cartographie).

Références

[1] Lalloué B, Monnez JM, Padilla C, Kihal W, Le Meur N, Zmirou-Navier D, Deguen S (2013). A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. *International Journal for Equity in Health*, **12**(1), 21.

PNN : une nouvelle bibliothèque R pour la modélisation d'un réseau de neurones probabilistes de Specht (Rencontres R, Lyon, le 27-28/06/2013)

P.-O. Chasset [a]

[a] Chercheur indépendant
Nancy, France
pierre-olivier@chasset.net

Mots clefs : Réseau de neurones artificiels, Probabilité.

Dans le domaine de l'apprentissage automatique, l'algorithme proposé par Specht [1] présente un intérêt important. Celui-ci n'a pourtant pas encore fait l'objet d'une implémentation dans le langage de programmation statistique R. La nouvelle bibliothèque logicielle *PNN* comble cette lacune.

Considérons un ensemble d'observations représentées par des variables quantitatives réelles. Nous réalisons une classification de ces observations en plusieurs groupes. Connaissant cet ensemble d'observations et le groupe associé à chacune d'elles, nous voulons prédire le groupe d'appartenance d'une nouvelle observation. Hastie *et al.* [2] exposent plusieurs méthodes pour résoudre ce problème d'apprentissage automatique, ou plus précisément un problème d'apprentissage supervisé, car un ensemble d'observations dont le groupe est connu a été sélectionné par un superviseur. Le réseau de neurones artificiels constitue une des méthodes. Fondée sur une analogie avec le réseau que forment les neurones du cerveau, cette méthode s'est montrée particulièrement adéquate pour toute une série de problèmes dont l'aide à la décision, la reconnaissance de régularités et la classification. De la même manière que le cerveau adapte sa structure en fonction des apprentissages, le réseau de neurones artificiels nécessite une phase préalable d'apprentissage visant à adapter ses paramètres en fonction des observations sélectionnées par le superviseur. La technique d'adaptation des paramètres la plus commune est la rétropropagation du gradient. Bien que les réseaux de neurones artificiels constituent une méthode statistique d'excellence, cette technique d'adaptation souffre cependant de la nécessité d'un nombre important d'observations dont le groupe est connu et, surtout, d'un temps de calcul très important. Specht [1] résout le problème en proposant un modèle de réseau de neurones appelé « *Probabilistic neural network* » ou *PNN*, permettant un apprentissage instantané et fonctionnant même avec un faible nombre d'observations.

Le réseau de neurones de Specht [1] est conçu selon quatre couches de neurones. Seuls les neurones de deux couches adjacentes sont interconnectés. L'information transite dans un seul sens, d'une couche n à une couche $n+1$. Chaque neurone d'une couche est dédié à une même tâche. La première couche associe à chaque neurone une variable de l'observation nouvelle. Ses informations sont distribuées à tous les neurones de la seconde couche. Dans cette couche, il existe un neurone par observation apprise. Chaque neurone calcule une distance euclidienne entre l'observation nouvelle et l'observation apprise, pondérée par un paramètre de lissage. Ce paramètre permet de contrôler la finesse de généralisation de la méthode. Il évolue sur l'ensemble des réels positifs non nuls et tend vers 0 lorsque le nombre d'observations d'apprentissage tend vers l'infini. Sur cette distance est appliquée ensuite une fonction d'activation exponentielle. Ces informations sont ensuite transférées à un neurone spécifique à un groupe d'observations, de la troisième couche, qui les somme. Une quatrième et dernière couche de neurones reçoit toutes les informations spécifiques à chaque groupe et opère la prédiction de la classe.

La description de ce fonctionnement s'apparente à un réseau de neurones artificiels classique. Elle en diffère cependant sur deux éléments. Premièrement, au lieu d'avoir un nombre réduit de neurones dans la deuxième couche, la méthode utilise un neurone pour chaque observation sélectionnée pour l'apprentissage, conservant ainsi la totalité de l'information initiale. Deuxièmement, la transformation utilisée à la fin de la seconde étape est une fonction d'activation exponentielle, au lieu d'une fonction sigmoïde couramment utilisée. La particularité supplémentaire de ce réseau de neurones réside dans son fondement probabiliste. Pour une nouvelle observation donnée, le réseau de neurones, au lieu de prédire uniquement son groupe d'appartenance, estime également ses probabilités d'appartenance à chaque groupe.

Ces particularités nous permettent d'accéder à un certain nombre d'avantages. Ainsi, en effectuant la prédiction directement avec les observations sélectionnées pour l'apprentissage, l'avantage de la méthode réside dans sa capacité d'apprentissage immédiate à partir d'un faible nombre d'observations. De plus, la méthode possède une faible complexité : un seul paramètre de lissage est à calibrer. Enfin, cette méthode permet la prise en compte de la connaissance acquise préalablement en ajustant les résultats par des probabilités *a priori*.

En revanche, dans le cas d'un nombre important d'observations d'apprentissage, l'inconvénient de la méthode est son temps de calcul pour réaliser une prédiction. Il sera plus long que les autres méthodes du fait de la nécessité pour chaque prédiction d'effectuer un calcul sur l'ensemble des observations ayant servies à l'apprentissage.

Au regard des avantages et du faible nombre d'inconvénients procurés par la méthode, nous avons réalisé une implémentation de celle-ci sous le logiciel statistique R. L'installation de cette bibliothèque exporte quatre fonctions : *learn* effectue l'apprentissage à partir d'une ou plusieurs observations avec une classe connue, *smooth* détermine le paramètre optimal de lissage, *perf* calcule la performance de la méthode, et *guess* permet d'estimer la classe d'une observation, ainsi que les probabilités d'appartenance à chaque classe. Chaque fonction a fait l'objet d'un contrôle qualité : une suite de tests de fonctionnalité vérifie le bon comportement des fonctions au cours du développement et lors de l'installation. L'usage de cette bibliothèque dans sa version 1.0.1 est facilité par la mise à disposition d'un jeu de données *norms*, d'un guide utilisateur avec des exemples [3] et d'un *post* [4] relatant les différentes ressources, dont quatre tutoriels sur l'installation, l'utilisation, l'optimisation et l'évaluation de la performance d'un réseau de neurones probabilistes de Specht.

Le nouveau programme *PNN* écrit dans le langage statistique R vient ainsi compléter l'imposante bibliothèque communautaire dans le domaine de l'apprentissage supervisé avec l'implémentation d'un réseau de neurones de Specht [1]. Cette nouvelle bibliothèque logicielle, utilisable sans connaissance particulière d'optimisation ou de calibrage, incorpore toutes les méthodes nécessaires permettant une prédiction immédiate de la classe d'une nouvelle observation même à partir d'un faible nombre d'observations d'apprentissage.

Références

[1] Specht, D. F. (1990) Probabilistic neural networks. *Neural networks*, 3(1):109–118.
[2] Hastie ,T., Tibshirani, R., Friedman, J. (2008) The Elements of Statistical Learning. Data-mining, Inference, and Prediction. Springer series in statistics. Springer, Berlin, 2˝édition.
[3] http://cran.r-project.org/web/packages/pnn/pnn.pdf
[4] Chasset P.-O. (2013). *PNN: Probabilistic neural network for the R statistical language*. Software, http://flow.chasset.net/r-pnn/

Estimation des risques dans les formes familiales de cancer.

Y. Drouet[a], **V. Bonadona**[a,b] and **C. Lasset**[a,b]

[a]Unité de Prévention et d'Epidémiologie Génétique
Centre Léon Bérard
28 rue laënnec - 69008 LYON

[b]Departement Biomaths-Santé, Equipe Epidémiologie et Santé Publique
UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive
Université Claude Bernard Lyon 1
43 bd du 11 novembre 1918 - 69622 VILLEURBANNE cedex

e-mail : youenn.drouet@lyon.unicancer.fr

Mots clefs : Statistique en génétique, Données familiales, Risque de cancer.

L'unité de Prévention et d'Epidémiologie Génétique du Centre Léon Bérard a pour mission principale de prendre en charge les familles à risque héréditaire de cancer. Cette activité nécessite d'estimer les risques de cancers des personnes suivies, et contribue ainsi au développement de nouvelles méthodes en statistique génétique. Nous présentons l'utilisation de R dans ce contexte. En utilisant les packages `kinship2` [1] et `BayesMendel` [2], nous montrons comment il est possible de calculer avec R : (i) la probabilité d'être porteur d'une mutation sachant l'histoire familiale, et (ii) le risque de cancer d'un individu en fonction de son génotype.

Les limites de ces outils sont cependant rapidement atteintes. A ce jour en effet, il n'est pas possible d'estimer le risque individualisé, en l'absence de mutation identifiée, d'un individu ayant une histoire familiale évocatrice d'un risque héréditaire [3]. Cette situation concerne pourtant la majorité des individus suivis dans les centres d'oncogénétique. D'autre part, il n'existe aucune fonction R permettant le calcul et l'optimisation des vraisemblances conditionnelles, prospectives et rétrospectives dans le cas de données familiales [4].

En s'appuyant sur une partie du code en C++ développé par B. Bonaiti dans le progiciel `GENERISK`, nous avons développé un prototype de fonction R permettant le calcul récursif de la vraisemblance d'une famille par l'algorithme d'Elston-Stewart [5]. Cette fonction utilise du code écrit en FORTRAN, ce qui permet de réaliser ce calcul connu pour sa complexité dans un temps très court, environ 200 fois plus rapidement que le même code écrit en R. L'interface en R permet d'utiliser : (i) la fonction d'optimisation `nlminb` pour obtenir les estimations des paramètres de la fonction de risque qui maximisent la vraisemblance (ii) le package `bootstrap` pour calculer les intervalles de confiance, et (iii) les possibilités graphiques avancées de R. A terme, cette fonction permettra d'estimer les risques de cancer même en l'absence de mutation identifiée, à l'instar de la méthode BOADICEA d'estimation du risque de cancer du sein [6]. Le calcul des vraisemblances conditionnelles, prospectives et rétrospectives, sera également disponible, de même que la "Genotype-restricted likelihood", seule méthode capable de fournir des estimations sans biais dans le cas de familles sélectionnées à partir de critères cliniques et de tests génétiques [7]. Cette nouvelle fonction constituera ainsi un outil d'aide à la décision pour le clinicien, et pourra également être utilisée librement par les chercheurs méthodologistes en statistique génétique.

Références

[1] Therneau T., Atkinson E., Sinnwell J., Matsumoto M., Schaid D. and McDonnell S. (2012). kinship2: Pedigree functions. R package version 1.3.7.

[2] Parmigiani G., Chen S., Wang W., Katki H., Blackford A. Adapted from C code written by Omar Aguilar, Giovanni Parmigiani. (2012). BayesMendel: Determining Carrier Probabilities for Cancer Susceptibility Genes. R package version 2.0-7.

[3] Drouet Y. (2012). Modélisation de la susceptibilité génétique non observée d'un individu à partir de son histoire familiale de cancer. Applications aux études d'identification pangénomiques et à l'estimation du risque de cancer dans le syndrome de Lynch.*Thèse de doctorat*, Université Claude Bernard Lyon 1.

[4] P. Kraft et D. C. Thomas. (2000). Bias and efficiency in family-based genecharacterization studies : conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet.* **66** :1119-1131.

[5] Elston RC, Stewart J. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered.* **21**:523-542.

[6] Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tryggvadottir L, Syrjakoski K, Kallioniemi OP, Eerola H, Nevanlinna H, Pharoah PD, Easton DF. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer.* **98**:1457-66

[7] Bonaïti B, Bonadona V, Perdry H, Andrieu N, Bonaïti-Pellié C. (2011) Estimating penetrance from multiple case families with predisposing mutations: extension of the 'genotype-restricted likelihood' (GRL) method. *Eur J Hum Genet.* **19**:173-9

Prévision de consommation électrique avec R

R. Nedellec[a]

[a]Département OSIRIS
EDF R&D
1, avenue du général de Gaulle
92140 Clamart, France
raphael.nedellec@edf.fr

Mots clefs : Consommation électrique, Modélisation, Generalized Additive Models.

Le groupe Prévision de consommation du département OSIRIS d'EDF R&D a pour mission l'élaboration d'algorithmes et de méthodes de prévision de consommation électrique. Ces prévisions seront utilisées dans diverses entités d'EDF pour répondre aux différents besoins métiers. Des modèles de prévisions à différents horizons de temps (infra-journalier, journalier, à horizon mensuel ou annuel), et à différentes échelles (maille nationale, agrégats de clients, ou à des mailles plus locales) ont ainsi été mis en place et sont utilisés aujourd'hui au sein d'EDF. La consommation électrique est très liée à différentes variables explicatives : des variables météorologiques, des variables calendaires, des variables économiques, etc. Depuis quelques années, des travaux ont été menés sur la prévision de consommation électrique par modèles GAM (Generalized Additive Models) [1], pour mettre en évidence ces relations souvent non linéaires [2]. Le package `mgcv` [3] nous permet d'utiliser des modèles de régression par splines pénalisées. Nous présenterons les avantages de cette modélisation :

- Approche semi-paramétrique pour capter les différentes relations non linéaires entre la consommation électrique et ses variables explicatives.
- Le package `mgcv` offre de nombreuses possibilités : pénalisation, choix des bases de splines, sélection automatique de variable.
- On retrouve en sortie des effets estimés assez facilement interprétables (avantage des modèles additifs).

De plus, nous verrons dans ce "lightning talk" comment l'utilisation de R et sa souplesse nous permettent d'utiliser ce type de modélisation de façon efficace sur une grande quantité de données, notamment en couplant les modèles GAM à du calcul massivement parallèle par le biais de packages tels que `doMC` [4].

Références

[1] Hastie, T.J., Tibshirani, R.J. (1990). Generalized Additive Models. *Chapman and Hall/CRC*.
[2] Pierrot, A., Goude, Y. (2011). Short-Term Electricity Load Forecasting With Generalized Additive Models. *Proceedings of ISAP power*, pp 593-600.
[3] Wood, S.N. (2006). Generalized Additive Models: An Introduction with R. *Chapman and Hall/CRC*.
[4] Revolution Analytics (2012). doMC: Foreach parallel adaptor for the multicore package. R package version 1.2.5. *http://CRAN.R-project.org/package=doMC*.

`autoplot` : ready-made plots with ggplot2

Jean-Olivier Irisson

Laboratoire d'Océanographie de Villefranche (LOV)
Station Zoologique, 181 Chemin du Lazaret
06230 Villefranche-sur-Mer
irisson@normalesup.org

Keywords : Visualisation, Versatility, Multivariate data analysis

Generic functions with methods dedicated to specific data types are one of the great strengths of R. Neophytes can just use `plot` or `summary` on almost any kind of objects and something clever and appropriate happens. For example, "plotting" a linear model object provides the usual diagnostic plots: residuals vs. fitted, quantile-quantile plot, etc. Advanced users can program methods for their own data types and the rest of their code is cleaner because of it (no more convoluted function names!).

Many packages provide `plot` methods for the class of objects they implement. Those usually involve some computation with the data from the provided object and calls to base graphical functions. For the quantile-quantile plot defined in `plot.lm`, appropriate quantiles of the normal distribution are computed based on the number of data points in the model and are then plotted against the residuals of said model. This process results in a ready-to-use plot but provides only limited flexibility through a few arguments or the usual graphical parameters (set by `par`).

The package `ggplot2` provides a *grammar* to define a graphic. Plots are constructed by assembling building blocks and defining how each column of data should be represented on the plot (as different colours, different shapes, etc.). The intricacies such as colour choice, legends, axes ranges are dealt with automatically. This paves the way for greater flexibility in the definition and usage of graphical functions.

To fill the gap between `plot` methods defined using base graphics and the versatility (and good looks!) of `ggplot2`, `autoplot` is a generic for which methods that output `ggplot` objects can be defined. Once those methods are defined, anyone can "just use `autoplot`" and something appropriate happens, auto-magically.

But `autoplot` goes beyond what is already possible through base graphics, because it allows the user to alter the plot following the usual grammar of `ggplot2`. The author of the `autoplot` method implements the preliminary computation and the skeleton of the plot, then the user can override those choices to tailor the plot to his/her needs. Furthermore, because the computation and actual plotting are split between two functions (`fortify` for the computation and `autoplot` for the plot) it is even possible for the user to implement a completely different plot based on the usual diagnostic variables extracted from an object.

The package `autoplot` (`https://github.com/jiho/autoplot`) aims at providing `fortify` and `autoplot` methods for many kinds of objects. It started with objects resulting from multivariate

data analysis methods (Principal Component Analysis, Correspondence Analysis, etc.). These examples will be used to illustrate the versatility of the approach and how it can help gain better insight into the data. They will also demonstrate how the separation between computation (`fortify`) and plotting (`autoplot`) gives the opportunity to unify the output of many functions implementing the same analysis in various packages, at little coding cost.

Des analyses exploratoires multidimensionnelles pour prédire la progression des patients en thérapie

T. Delespierre [a,b], Céline Piedvache [b], Jean-Michel Thurin [a], Monique Thurin [a], B. Falissard [a]

[a] Unité INSERM 669
77 bd de Port Royal, 75769 Paris, France
tiba.baroukh@gmail.com
jmthurin@internet-medical.com
mthurin@internet-medical.com
falissard_b@wanadoo.fr

[b] Unité de Recherche Clinique
CHU Bicêtre, 94275 Le Kremlin-Bicêtre, France
celine.piedvache@gmail.com

Mots clefs : Analyse en Composantes Principales, ACP, Classification Ascendante Hiérarchique, CAH, RRFPP, psychothérapie, autisme, ECAR, EPCA, CPQ.

Résumé
Après 2 mois de suivi, la progression de 50 patients autistes en thérapie dans le cadre du Réseau de Recherches Fondé sur les Pratiques Psychothérapiques (RRFPP [1]) est prévisible statistiquement à l'aide d'outils d'analyse exploratoire multidimensionnelle (ACP et CAH).

Contexte
Le projet RFPP a été sélectionné en février 2008 dans le cadre de l'appel à projets INSERM 2007 'Réseaux de recherche clinique et en santé des populations' conçu pour répondre au manque d'évaluation des pratiques psychothérapiques en France.

Une centaine de cas suivis en psychothérapie ont été progressivement inclus à compter de 2008 et ont fait l'objet, dans un premier temps, d'une étude systématique processus-résultats multidimensionnelle durant 1 an, puis d'une analyse comparative des différences et des communautés entre cas analogues à partir de leur réunion dans une base de données.

Patients et Thérapeutes
Les cliniciens travaillent en groupes de pairs et pratiquent l'intervision (3 psychiatres et /ou psychologues cliniciens exerçant en institution ou praticiens en ville). Le réseau suit principalement des patients borderline (adolescents et adultes) et autistes (enfants).

Chaque cas démarre avec les notes extensives du psychothérapeute des 3 premiers entretiens, puis au cours du temps, à 2 mois, 6 mois et 12 mois. L'analyse de ces documents permet de générer une évaluation qualitative de la formulation de cas du patient et de définir les modérateurs (qualité du plateau technique, comorbidités psychiques et physiologiques, soutien familial, scolarisation).

La formulation de cas est suivie d'une évaluation quantitative des changements au travers de trois instruments validés et hétéroévalués par le groupe de pairs:
- deux outils d'évaluation de l'état de santé : l'ECAR [2], Echelle des Comportements Autistiques Révisée (29 items cotés de 0 à 4 et 2 facteurs); l'EPCA [3], Evaluation

Psychodynamique des Changements Autistiques (140 items cotés de 0 à 3, 5 niveaux de développement et 8 dimensions)
- un outil d'analyse de la séance de thérapie, le CPQ [4], Child Psychotherapy Q-set (100 items cotés de -4 à +4 dont les scores suivent une loi normale par blocs, suivant la méthodologie du Q-sort). Le CPQ mesure la dynamique de la diade patient-thérapeute.

Méthode
Chaque patient est décrit sur la période inclusion-2 mois: âge, sexe, nombre d'années en thérapie, modérateurs (9 variables binaires alimentées par lecture de la formulation de cas), scores ECAR et EPCA à baseline et à 2 mois, scores CPQ à 2 mois.
Une sélection des variables normalisées pertinentes (CPQ exclu), est faite à l'aide de la fonction sphpca de la library psy : la visualisation des variables sur la sphère permet d'exclure l'information redondante. Une analyse en composantes principales à l'aide de la fonction prcomp du package FactoMineR confirme cette sélection. Suit alors une classification ascendante hiérarchique (fonction hclust du package cluster) à partir de la matrice de distance maximum des 50 patients décrits par les variables obtenues à l'étape précédente et celles du CPQ à 2 mois. La visualisation du dendrogramme permet de sélectionner une partition en 4 classes (fonction cutree du package cluster) avec un bon gradient d'inertie selon le critère de Ward.

Résultats
La recherche d'un sens 'naturel' des classes générées aboutit au regroupement des classes 3 et 4 qui évolueront plus lentement que le reste de la population (classes 1 et 2 regroupées), en termes de gravité ECAR, de développement EPCA et du nombre d'aptitudes acquises (sous échelle de l'EPCA). Les deux ensembles de patients sont comparés à 12 mois. Avec ou sans condition de normalité (t.test ou wilcox.test), les résultats concordent :
- diminution de la gravité ECAR : 15.6 versus 27.2 p-value=2.5 e-04
- augmentation du développement EPCA : 59.5 versus 34.6 p-value=5.6 e-04
- nombre d'aptitudes acquises : 11 versus 5.5 p-value=2.8 e-05

Discussion
Tous les outils utilisés (ECAR, EPCA et CPQ) ont été validés. Variables quantitatives et qualitatives (sexe, indicatrices d'état, modérateurs) ont été intégrées dans un même modèle moyennant un peu de data management. Les deux groupes obtenus à partir des données des 2 premiers mois sont significativement différents. Ces résultats ont été obtenus a posteriori, mais avec les futures inclusions il sera possible de vérifier ces hypothèses a priori.

Références
[1] JM Thurin, M Thurin. Réseau de Recherches Fondées sur les Pratiques Psychothérapiques : le pôle autisme. Pour la recherche 2011; 68-69: 15-16.
[2] ECA-R Barthélémy et al, 1997.
[3] http://www.genevievehaagpublications.fr/
[4] C Schneider The Development of the CPQ-Set University of California, Berkeley, 2004
[5] T Baroukh Combler le fossé entre pratique clinique et recherche Journées Fouille de Données Complexes et de Grands Graphes. CNAM 20-21 juin 2011
[6] T Delespierre New Tools for studying psychotherapies Rencontres R-Bordeaux 2012
[7] T Delespierre, JMThurin, M Thurin, B Falissard. Aggregating cases IACAPAP 2012

R au secours des écotoxicologues

G. Kon Kam King[a], **P. Veber**[a], **M.L. Delignette-Muller**[a,b] and **S. Charles**[a,c]

[a]Université de Lyon; Université Lyon 1
Laboratoire de Biométrie - Biologie Evolutive
CNRS; UMR 5558
43 boulevard du 11 novembre 1918
69622 Villeurbanne Cedex
guillaume.kon-kam-king@univ-lyon1.fr
philippe.veber@univ-lyon1.fr

[b]Université de Lyon
VetAgro Sup Campus Vétérinaire de Lyon
1 avenue Bourgelat
69280 Marcy l'Etoile
marielaure.delignettemuller@vetagro-sup.fr

[c]Institut Universitaire de France
103 boulevard Saint-Michel
75005 Paris
sandrine.charles@univ-lyon1.fr

Mots clefs : interfaçage web, analyse de données de bioessais, distribution de sensibilité des espèces.

Le défi majeur que doit relever aujourd'hui l'écotoxicologie est de se munir d'outils de modélisation prédictifs, intégrés dans un cadre décisionnel standardisé, et dont les autorités de régulation et les décideurs puissent directement tirer profit pour une meilleure gestion des impacts potentiels des substances chimiques sur les populations et les communautés à protéger. Dans cette perspective, il apparaît nécessaire d'offrir aux écotoxicologues le moyen d'exploiter au mieux les données qu'ils acquièrent. Les données en écotoxicologie sont issues de bioessais, généralement standardisés selon des normes (OCDE, ISO), en toxicité aiguë ou chronique, au cours desquels sont mesurées la survie, la reproduction et/ou la croissance d'organismes modèles. Ce sont donc des données temporelles dépendantes d'une gamme croissante de concentrations testées.

Classiquement, l'analyse statistique de ce type de données se réduit à l'estimation de concentrations critiques d'effet à un temps donné (le plus souvent en fin d'essai), soit en utilisant des tests d'hypothèse (par exemple pour déterminer une NOEC, *No Observed Effect Concentration*), soit en ajustant des modèles paramétriques concentration-réponse (par exemple pour estimer des ECx, *x% Effective Concentration*). Malgré l'existence d'un guide OCDE [1], la jungle des statistiques paraît généralement impénétrable à l'écotoxicologue : il gaspille une bonne partie de ses données à ne les considérer qu'en fin d'essai, il doit choisir parmi une multitude de tests/méthodes/modèles celui qui "convient le mieux", la nature de ses données ne lui permet pas toujours de vérifier les conditions d'applications de ces tests/méthodes/modèles, il ne dispose pas d'un outil "clé en main", fiable statistiquement, convivial et simple d'utilisation...

Face aux macros Excel faites maison ou aux logiciels boîte noire parfois hors de prix, R pourrait être l'alternative salvatrice. Pour des analyses statistiques simples, il existe déjà quelques packages dédiés comme par exemple `drc` ou `fitdistrplus`. Le package `drc` permet l'estimation des paramètres des modèles concentration-réponse par maximum de vraisemblance [2] ; le package `fitdistrplus` permet d'ajuster des distributions paramétriques par maximum de vraisemblance, le cas échéant en tenant compte de la présence de données censurées [3]. Par contre, pour des analyses plus complexes, par exemple avec des données de comptage qui nécessitent d'utiliser des modèles d'erreur non standard ou avec des données dépendantes du temps, il faut des programmes R plus sophistiqués. Et dans l'optique que ces programmes R soient utilisés par des écotoxicologues, que les aspects techniques et théoriques sous-jacents rebutent généralement, il est indispensable d'investir dans la conception d'interfaces utilisateur faciles d'utilisation. L'accessibilité d'un tel outil repose sur deux aspects : (1) l'interface graphique rend transparente l'utilisation plus experte de l'interprêteur R ; (2) elle propose un paramétrage par défaut ou automatique des procédures d'estimation.

Dans cet exposé, nous commencerons par illustrer la valeur ajoutée de R pour des analyses statistiques simples de données d'écotoxicologie. Nous dévoilerons ensuite notre nouvelle interface, MOSAIC_SSD (*Modelling and Statistical Tools for Ecotoxicology*, `http://pbil.univ-lyon1.fr/software/mosaic/`), qui permet de modéliser de façon simple et conviviale la distribution de sensibilité d'une communauté d'espèces à une substance chimique donnée, pour en extraire par exemple une concentration qui affecte $x\%$ (ou qui protège 1-$x\%$) des espèces considérées (HCx, $x\%$ Hasard Concentration). Nous montrerons également comment cette interface pallie les manques actuels des logiciels existants, en permettant d'une part de calculer des intervalles de confiance bootstrap autour des HCx, et d'autre part de prendre en compte des données censurées. Enfin, nous discuterons des spécificités inhérentes à la conception de ce type d'interface. Nous évoquerons l'approche suivie pour embarquer un interpréteur R dans un serveur web, en particulier les aspects relatifs au parallélisme et à l'appel de fonctions R depuis le langage utilisé par le serveur.

Références

[1] OCDE (2006) Current approaches in the statistical analysis of ecotoxicity data: a guidance to application. Technical report, Organisation for Economic Cooperation and Development.
[2] Ritz, C., & Streibig, J. (2005). Bioassay analysis using R. Journal of Statistical Software, 12(5), 1–22.
[3] Delignette-Muller ML, Pouillot R, Denis JB, Dutang C. (2009). `fitdistrplus`: Help to Fit of a Parametric Distribution to Non-censored or Censored Data. Available at `http://cran.at.r-project.org/web/packages/fitdistrplus/index.html` [accessed 7 May 2013].

R-chaeology
Analyse et datation d'artefacts archéolgiques: R et les cachets circulaires hittites

N. Strupler [a] [b]

[a] UMR 7044 « Archéologie et Histoire ancienne : Méditerranée-Europe » (ARCHIMÈDE)
Université de Strasbourg
5 allée du Gal Rouvillois, 67083 Strasbourg

[b]Institut für Altorientalische Philologie und Vorderasiatische Altertumskunde
Westfälische Wilhelms-Universität Münster
Rosenstr. 9, 48143 Münster, Allemagne

nehemie.strupler@etu.unistra.fr

Mots clefs : statistique, chronologie, archéologie

Dans cette présentation nous montrerons comment les analyses statistiques permettent aux sciences historiques, notamment à l'archéologie, de renouveler des questionnements à travers une étude de cas. Les outils statistiques sont encore peu employés par les archéologues et les philologues : des corpus immenses existent mais ne donnent généralement lieu à aucune autre étude statistique qu'un simple dénombrement. Nous traiterons des cachets circulaires hittites et nous regarderons comment une analyse factorielle permet d'affiner la typologie et la datation d'artefacts. Une fois l'évolution des objets mieux définie il est possible d'inférer sur la signification des changements au sein de la civilisation. Après une présentation du corpus et de la méthode d'investigation, les résultats illustreront des perspectives offertes par R aux archéologues et aux chercheurs en sciences historiques.

Utilisés depuis des millénaires, les cachets de pierre ou de métal gravés en creux ou en relief d'initiales, de titres ou d'emblèmes servent à sceller des objets pour en attester l'authenticité et contrôler leur ouverture. Les cachets et leurs empreintes témoignent des pratiques aussi bien administratives qu'artistiques de la civilisation qui les a produits. Le corpus des cachets hittites offre l'avantage d'être assez large et homogène pour être étudié statistiquement.
De la civilisation hittite (Anatolie centrale, ca 1650–1200 av. n. è.) nous connaissons un corpus important d'empreintes de cachets circulaires; plus de 3000 ont été découvertes dans la capitale Hattuša, aujourd'hui à environ 150 km à l'est d'Ankara [1–3, 5–6]. Ces cachets appartiennent aux rois, aux reines ou à certains fonctionnaires. Au sein des cette production, on connait bien l'évolution des sceaux royaux dont l'ordonnance chronologique est assurée. En revanche, l'évolution des sceaux des fonctionnaires est moins bien connue.
Tous les cachets circulaires ont été réalisés selon le même modèle avec une plage centrale et des bordures circulaires (fig. 1). La plage centrale se compose généralement d'une inscription en louvite hiéroglyphique mentionnant le titre et le nom du propriétaire accompagnée de divers motifs. Les bordures circulaires peuvent être garnies d'une légende en akkadien, c'est le cas pour les cachets royaux, de hiéroglyphes, de représentations humaines ou bien de motifs décoratifs. Parmi certaines variables retenues dans cette étude, le motif central, la taille de la plage centrale, des bordures circulaires, la forme du cachet, les hiéroglyphes, les éléments décoratifs récurrents (triangles, cercles, motifs trilobés, coniques ou encore aigles héraldiques bicéphales) permettent

Figure 1: Deux empreintes de cachets circulaires hittites (ca. 14$^{\text{ème}}$ siècle av. n. è.) D'après [5] Cat. 172 et 178.

d'obtenir des données quantitatives et qualitatives.
À l'aide d'une analyse factorielle des données mixtes et du package FactoMineR [7, 8] nous étudierons la variabilité des cachets cylindriques. Tout d'abord, avec une analyse en composante principale des données métriques, il est possible de distinguer différents groupes. Les hiéroglyphes seront étudiés grâce à une analyse factorielle des correspondances. Grâce à la modélisation des paramètres quantitatifs et qualitatifs, nous illustrerons une typologie différente de celle utilisée actuellement [10]. Les artefacts bien datés à travers leur contexte archéologique ou philologique serviront de piliers pour transcrire la variabilité en une évolution chronologique. Finalement nous soulignerons les apports de l'analyse factorielle et son accessibilité grâce à R : la corrélation entre la datation et les données métriques des cachets, le choix des hiéroglyphes, le titre, la représentation iconographique et les éléments décoratifs offrent une nouvelle base pour interpréter les cachets dans leur contexte historique.

Références

[1] Boehmer, R. M. ; H. G. Güterbock (1987). *Glyptik aus dem Stadtgebiet von Boğazköy*, Gebr. Mann, Berlin
[2] Güterbock, H. G. (1940). *Die Königssiegel der Grabungen bis 1938*, Weidner, Berlin
[3] Güterbock, H. G. (1942). *Die Königssiegel von 1939 und die übrigen Hieroglyphensiegel*, Weidner, Berlin
[4] Hawkins, J. D. (2000). *Corpus of hieroglyphic Luwian inscriptions*, de Gruyter, Berlin
[5] Herbordt, S. (2005). *Prinzen- und Beamtensiegel der hethitischen Grossreichszeit auf Tonbullen aus dem Nişantepe-Archiv in Hattuša*, von Zabern, Mayence
[6] Herbordt, S. ; D. Bawanypeck, ; J. D. Hawkins (2011). *Die Siegel der Grosskönige und Grossköniginnen auf Tonbullen aus dem Nişantepe-Archiv in Hattuša*, von Zabern, Mayence
[7] Husson, F. ; J. Josse ; S. Lê ; J. Mazet (2013). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.23
[8] Husson, F. ; S. Lê ; J. Pagès (2011). *Exploratory Multivariate Analysis by Example Using R*, CRC Press, Boca Raton
[9] Laroche, E. (1960). *Les hiéroglyphes hittites*, Éditions du CNRS, Paris
[10] Mora C. (1987). *La glittica anatolica del II millenio A.C.: Classificazione Tipologica I. I sigili a iscrizione geroglifica*, Gianni Iuculano, Pavie

Teaching R to social science undergraduates

F. Briatte[a] and **I. Petev**[b]

[a]UMR CNRS Pacte
Institut d'Études Politiques de Grenoble
19 rue Pajol, 75018 Paris
f.briatte@ed.ac.uk

[b]Observatoire Sociologique du Changement
Sciences Po
1, rue Stendhal, 75020 Paris
ivaylo.petev@ensae.fr

Keywords : Statistics, Teaching, Social Sciences.

This lightning talk is based on our recent experience teaching R (through RStudio) to a small group of undergraduates with social science backgrounds. It covers:

1. the course itself, its student audience, and how it all went;
2. what we think could be good teaching practices in that area; and
3. what we believe could help teach R to non-programmer audiences.

The talk will focus on three particular aspects of our teaching experience:

- the importance of teaching applied, empirical examples to enhance interest and understanding [1], as well as student ability to link statistics and real-world situations [2]
- the importance of teaching data retrieval and management skills along the common curriculum of introductory statistics
- challenges posed by the R software environment to absolute beginners, such as interpreting error output and reproducibility at scale

Our small-scale teaching experiment intends to show that R can be effectively taught to audiences whose specialized skills lie outside of statistical computing.

All material for our course and this talk will be made available through GitHub in replicable formats, along with examples of student work. Our syllabus is attached to this abstract.

Note: our course was taught in English, and we offer to write our talk in English, but we both speak French and are naturally willing to use French for any or all parts of our talk.

References

[1] Genolini, C., Driss, T. (2010). Éveiller l'intérêt pour la statistique par l'exemple. *Statistique et Enseignement*, **1**(2), 49-57

[2] Yilmaz, M. R. (1996). The Challenge of Teaching Statistics to Non-Specialists. *Journal of Statistics Education*, **4**(1): `http://www.amstat.org/publications/jse/v4n1/yilmaz.html`

R tools for spatial point pattern analysis applied to fluorescence localization nanoscopy

J. Godet[a,b], **H. Anton**[a] and **Y. Mély**[b]

[a]Laboratoire de Biophotonique et Pharmacologie
Université de Strasbourg, UMR CNRS 7213
Faculté de Pharmacie, 74 rte du Rhin, Illkirch, France
julien.godet@unistra.fr

[b]Département d'Information Médicale et de Biostatistiques
Hôpitaux Universitaires de Strasbourg
1, pl de l'Hôpital, Strasbourg, France

Mots clefs : Nanoscopy, Spatial Statistics, R.

Emerging super-resolution optical microscopy techniques (usually referred as nanoscopy) capable of operating beyond the diffraction limit give now access to cell images with unprecedented levels of details [1-3]. These breakthrough technologies are particularly suitable to study the localization of fluorescent nano-objects within the cell environment [4]. Localization nanoscopy can image biological samples with high molecular densities while maintaining the localization accuracy of single nano-particles. But if localization nanoscopy can be routinely performed on conventional fluorescence microscopes, the challenge is now to offer data analysis facilities allowing a straightforward translation from single molecules detection to biological insights.
Here we propose few application examples using the point pattern analysis tools developed in R [5] to highlight their ability to extract valuable and biologically relevant information on nanoparticles distribution in the intracellular organelles. Taken together, coupling latest imaging techniques and R data-analysis facilities holds the promise to go one step further in the understanding of biological structures and dynamics.

References

[1] Beyond the diffraction limit (2009). *Nature Photonics - Special Issue*, **3**(7)
[2] Rust, M.J., Bates, M., Zhuang X. (2006). Stochastic optical reconstruction microscopy (STORM) provides sub-diffraction-limit image resolution. *Nature Methods*, **3**(10), 793-795.
[3] van de Linde S., Löschberger A , Klein T., Heidbreder M., Wolter S., Heilemann M., Sauer M. (2011). Direct stochastic optical reconstruction microscopy with standard fluorescent probes *Nature Protocols*, **6**, 991-1009
[4]Sharma P., Brown S., Walter G, Santra S., Moudgila B. (2006). Nanoparticles for bioimaging *Advances in Colloid and Interface Science*, **123**(126), 471-485.
[5] Baddeley A. and Turner R. (2005). Spatstat: an R package for analyzing spatial point patterns *Journal of Statistical Software* **12**(6), 1-42.

Traiter des données de tracking avec R
navigation sur un site web, suivi d'enquête web ou téléphonique

Anne GAYET

Directrice Datamining
A.I.D.
4 rue Henri le Sidaner
78000 Versailles
agayet@aid.fr

Mots clefs : séquences d'événements, règles d'association, motifs séquentiels, co-clustering, trajectoires, données de tracking, webmining. Packages arules, arules sequences, TraMineR, isa2.

Les suites d'évènements que constituent les données de tracking nécessitent des méthodologies particulières pour être traitées. Ces données sont composées d'états ordonnés dans le temps, éventuellement associés à des durées. Par exemple :

- pour le suivi d'enquête téléphonique : les états sont les différentes tentatives d'appel,
- pour le suivi d'enquête web : les états sont les écrans affichant une ou plusieurs questions, les écrans sont horodatés lors de leur affichage,
- en webmining : les états sont les pages vues lors d'une visite sur un site web et sont horodatées.

Pour synthétiser et visualiser ces types de données, les méthodes traditionnelles d'analyse de données multidimensionnelles peuvent certes être utilisées : création d'indicateurs agrégés comme le nombre d'états, la durée totale, etc … puis utilisation des méthodes d'analyse factorielle et / ou typologie. Toutefois elles ne prennent pas en compte la dimension temporelle des données, ni le besoin de traiter conjointement les individus (internautes, enquêtés) et les nombreux états suivis (pages vues, questions et items afférents). Des méthodes plus spécifiques le permettent : règles d'association, motifs séquentiels, clustering de séquences. Nous verrons comment mettre en œuvre ces méthodes avec les packages Arules, AruleSequences, et TraMineR.
Une autre façon de synthétiser ce type de données est d'utiliser le co-clustering (ou biclustering) sur la matrice des fréquences individus x nombre de passages par chaque état. Nous verrons comment le mettre en œuvre avec le package isa2.

Par ailleurs, la visualisation des données détaillées d'une part, des synthèses obtenues d'autre part, nécessite des modes de représentation très différentes des visualisations traditionnelles. Nous verrons comment visualiser les associations ou motifs séquentiels en utilisant des graphes de liens dans lesquels les nœuds sont des états et les liens sont les fréquences de passage entre les états.

Références

[1] Nicolas S. Müller, Matthias Studer, Alexis Gabadinho, Gilbert Ritschard (2010). Analyse de séquences d'événements avec TraMineR. EGC 2010.

[2] Alexandre Pollien (FORS), Dominique Joye (ISS), Michèle Ernst Stähli (FORS), Marlène Sapin (FORS) - Université de Lausanne (2012). Répondants et non-répondants dans les enquêtes, analyse des séquences de contact. 7ème colloque francophone sur les sondages, Rennes.

[3] Alexis Gabadinho, Gilbert Ritschard, Matthias Studer and Nicolas S. Muller (2011) Department of Econometrics and Laboratory of Demography - University of Geneva. Mining sequence data in R with the TraMineR package: A user's guide http://mephisto.unige.ch/traminer/

[4] Buchta, C., Hahsler, M. (2010). "arulesSequences: Mining frequent sequences". R package. http://CRAN.R-project.org/package=arulesSequences

Les interfaces graphiques conviviales, ou comment rendre R plus accessible aux utilisateurs non-informaticiens

J.P. Maalouf [a], S. Longis [b], S. Montaudouin [c], C. Vieuille [d], G. Le Pape [e]

AnaStats Scop ARL
Les Vigneaux, 37220, Rilly Sur Vienne, France
[a] jeanpaul.maalouf@anastats.fr
[b] sandrine.longis@anastats.fr
[c] severine.montaudouin@anastats.fr
[d] caroline.vieuille@anastats.fr
[e] lepape.gilles@anastats.fr

Mots clefs : convivialité, débutants, interfaces graphiques, non informaticiens, R Commander, synthèse

Une partie considérable des utilisateurs de R est constituée de non-informaticiens ayant souvent peu de temps à consacrer aux analyses statistiques. Aborder les analyses à travers des lignes de commande est rébarbatif pour ce public. En plus, plusieurs analyses exigent le recours à une succession d'un grand nombre de lignes de commande pour qu'elles soient complètes, ce qui rend les choses encore plus compliquées. Cette présentation a pour but de souligner l'importance des interfaces graphiques telles que R Commander pour une grande partie de la communauté scientifique utilisant R. En effet, notre expérience en tant que formateurs aux statistiques - via le logiciel R notamment - nous montre que, de par leur convivialité, ces interfaces sont très appréciées par les chercheurs : elles offrent la possibilité de manipuler des données et de faire des graphiques et des analyses plus ou moins poussées et synthétiques à travers des menus déroulants, sans recours à des lignes de commande. Nous espérons que ce type d'interfaces sera amené à être développé davantage par les informaticiens-statisticiens.

TraMineR : Une boite à outils pour l'exploration et la visualisation de séquences

Gilbert Ritschard

Pôle national de recherche LIVES
Institut d'études démographiques et du parcours de vie
Université de Genève, 40, bd du Pont d'Arve, CH-1211 Genève 4, Suisse
gilbert.ritschard@unige.ch

Mots clefs : Séquences d'états, séquences d'événements, visualisation, dissimilarités, analyse basée sur les dissimilarités.

TraMineR est une librairie dévolue à l'exploration de séquences, essentiellement de séquences d'états et d'événements ordonnés chronologiquement [2]. On rencontre de telles séquences dans des domaines très divers tels que le contrôle d'appareils où l'on examine les séquences d'états de fonctionnement, en gestion où l'on s'intéresse par exemple aux successions d'achats de clients ou d'activités exercées par des employés, en analyse de l'usage du web où l'on analyse des séquences de pages visitées, et en analyse des parcours de vie où l'on étudie des séquences décrivant notamment des carrières professionnelles ou des vies familiales. Certaines des fonctionnalités proposées par TraMineR s'appliquent également à des séquences non chronologiques comme les séquences de lettres ou mots en analyse de textes, ou encore les séquences de protéines ou de nucléotides en biologie, pour lesquelles d'autres outils s'avèrent cependant mieux adaptés (voir `http://www.bioconductor.org/`).
La librairie TraMineR a été conçue à l'origine pour répondre à des questions liées à l'analyse de parcours de vie où les données comprennent typiquement quelques centaines, voire quelques milliers de séquences de longueur comprise entre 10 et 100 lorsqu'il s'agit de séquences d'états et incluant rarement plus d'une dizaine d'événements dans le cas de séquences d'événements. L'alphabet des états ou événements compte le plus souvent moins de 15 ou 20 éléments.
TraMineR offre des outils pour explorer des séquences d'états aussi bien que des séquences d'événements datés. Les séquences d'états se caractérisent par le fait que la position dans la séquence porte une information temporelle, à savoir la durée depuis le début de l'observation (par exemple le nombre de mois après la fin de la scolarité obligatoire), tandis que dans les séquences d'événements, la date de chaque événement doit être explicitement attachée à chaque événement (marié à 25 ans, premier enfant à 27 ans).

Tableau 1 – Vue transversale (gauche) versus vue longitudinale (droite)

id	t_1	t_2	t_3	$\cdots$
1	JL	JL	EM	$\cdots$
2	SC	SC	TR	$\cdots$
3	SC	SC	SC	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

id	t_1	t_2	t_3	$\cdots$
1	JL	JL	EM	$\cdots$
2	SC	SC	TR	$\cdots$
3	SC	SC	SC	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Les séquences d'états peuvent être organisées sous la forme illustrée au tableau 1 où chaque ligne correspond à un cas et chaque colonne aux unités de temps. On peut changer l'alignement en passant par exemple d'un alignement sur les dates à un alignement sur l'âge (durée du processus). Les outils proposés pour les séquences d'états permettent en particulier de

– rendre compte de l'évolution des distributions transversales (chronogrammes, entropies transversales, ...) ;
– visualiser l'ensemble des séquences individuelles et de calculer des caractéristiques (nombre de changements d'états, durées moyennes dans les états ou complexité de la séquence par exemple) ;
– calculer la dissimilarité entre séquences selon plusieurs métriques ;
– visualiser des groupes et donc en particulier les clusters qui peuvent être déduits des dissimilarités.

La librairie offre également plusieurs outils d'analyse originaux fondés sur les dissimilarités :
– calculer et analyser la dispersion des séquences [6] ;
– identifier visualiser des séquences représentatives (medoid, avec plus forte densité, ...) [3] ;
– générer des arbres de régression de séquences [6].

Les séquences d'événements se distinguent des séquences d'états par l'absence d'alignement sur une date ou un âge et la possibilité d'avoir des événements simultanés. Les outils spécifiques offerts sont [4] :
– visualisation sous forme de 'parallel coordinate plot' [1] ;
– extraction de sous-séquences fréquentes sous diverses contraintes de temps, de contenu et de selon diverses méthodes de comptage ;
– identification des sous-séquences les plus discriminantes entre groupes ;
– calcul de dissimilarités entre séquences d'événements.

La librairie inclut également plusieurs fonctions utilitaires notamment pour convertir entre diverses possibilités d'organisation des données et en particulier pour aider à convertir entre séquences d'états et séquences d'événements datés [5].

La présentation portera sur la genèse de la librairie et sa philosophie axée sur des objets séquences d'états et séquences d'événements qui incluent un maximum d'information comme l'alphabet, les étiquettes courtes et longues, la palette de couleur pour les visualisation et les pondérations pour n'en citer que quelques uns. Nous évoquerons également l'attention accordée à la documentation et au support offert aux utilisateurs.

Références

[1] Bürgin, R. and G. Ritschard (2012). Categorical parallel coordinate plot. In *LaCOSA Lausanne Conference On Sequence Analysis, University of Lausanne, June 6th-8th 2012*, Lausanne. Poster.

[2] Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011a). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software* **40**(4), 1–37.

[3] Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2011b). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

[4] Ritschard, G., R. Bürgin, and M. Studer (2013). Exploratory mining of life event histories. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology. New York : Routledge. (in press)

[5] Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin : Springer-Verlag.

[6] Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research* **40**(3), 471–510.

multlcmm: fonction d'estimation de modèles mixtes à processus latent pour données longitudinales multivariées

V. Philipps[a,b] and **C. Proust-Lima**[a,b]

[a]INSERM U897
F-33076 Bordeaux, France

[b]Université Bordeaux Segalen
F-33076 Bordeaux, France
Viviane.Philipps@isped.u-bordeaux2.fr
Cecile.Proust@isped.u-bordeaux2.fr

Mots clefs : modèles mixtes multivariés, processus latent, données longitudinales.

En psychologie et en santé notamment, les variables d'intérêt sont souvent des concepts ou des quantités qui ne sont pas directement observables. C'est le cas notamment de la cognition, de la qualité de vie, de la douleur ou encore du bien-être. En pratique, ces quantités sont mesurées par un ensemble d'échelles ou de variables. La cognition est par exemple mesurée par une batterie de tests psychométriques et la qualité de vie est mesurée par un ensemble de questionnaires.

Dans les études transversales, ce type de données a donné lieu à de nombreux développements en modèles à équations structurelles ou modèles à variables latentes avec notamment la théorie des réponses aux items.

Dans les études longitudinales, la quantité sous-jacente devient un processus latent dont on cherche à décrire la trajectoire en fonction du temps. Nous proposons dans la fonction multlcmm du package lcmm de décrire son évolution par un modèle linéaire mixte et de relier ce processus latent aux différentes variables observées par des transformations nonlinéaires paramétrées. Ces modèles généralisent ainsi la théorie des modèles à effets aléatoires au cas multivarié et étendent les modèles à variables latentes au cas longitudinal [1,2].

Outre la possibilité d'inclure dans le modèle linéaire mixte un processus d'autocorrélation en plus des effets aléatoires, n'importe quelles fonctions du temps et n'importe quelles variables explicatives, cette approche permet aussi de traiter des données de natures différentes. En effet, l'équation d'observation définie pour relier les variables observées et le processus latent aux temps d'observations peut prendre différentes formes. De cette façon, des variables quantitatives Gaussiennes aussi bien que des données curvilinéaires (c'est-à-dire avec des effets plafonds et planchers et une sensibilité variable au changement comme c'est souvent le cas en psychologie) peuvent être traitées.

De plus, la restriction des modèles à variables latentes classiques aux données équilibrées est supprimée. Cette méthode permet d'analyser des données complètement déséquilibrées, c'est-à-dire avec des temps de mesure possiblement différents d'un sujet à l'autre et différents d'une variable à l'autre. Aussi, des classes latentes d'évolution peuvent être incorporées dans le modèle linéaire mixte sous-jacent, ce qui permet une analyse de profils de trajectoires pour le processus latent.

Comme dans les autres fonctions du package lcmm, les paramètres de ce modèle sont obtenus par maximum de vraisemblance à l'aide d'un algorithme de Marquardt [3] modifié avec des critères d'arrêt stricts (sur les dérivées premières et secondes) [2]. Plusieurs fonctionnalités sont implémentées à partir d'un objet multlcmm, notamment des calculs de prédictions, et des graphiques des fonctions de lien estimées.

L'objectif de la présentation est d'introduire ce type de modèle, d'en détailler l'implémentation dans multlcmm et d'en illustrer l'utilisation sur des données réelles de trajectoire cognitive chez les personnes âgées mesurées par divers tests psychométriques [4].

Références

[1] Proust-Lima, C., Amieva, H., Jacqmin-Gadda, H. (2012). Analysis of multivariate data : a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology.*
[2] Proust C., Jacqmin-Gadda H., Taylor J. M., Ganiayre J., Commenges D. (2006). A non-linear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, **62** (4), 1014-24.
[3] Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431-41.
[4] Proust-Lima, C., Amieva, H., Dartigues, J.-F., Jacqmin-Gadda, H. (2007). Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population-based studies. *American Journal of Epidemiology*, **165** (3), 344-50.

Prédiction d'un événement binaire à partir de données fonctionnelles : Application aux bovins laitiers

C. Sauder[a] and **H. Cardot**[b]

[a]UMR 1348 PEGASE
INRA-Agrocampus Ouest
65 rue de Saint-Brieuc, Rennes
cecile.sauder@rennes.inra.fr

[b]Institut de Mathématiques
Université de Bourgogne
9 av. Alain Savary, Dijon
herve.cardot@u-bourgogne.fr

Mots clefs : Prédiction, régression logistique, données fonctionnelles.

L'objectif de l'étude est de prévoir le succès (ou non) à la première insémination d'une vache à partir des courbes de lactation des 42 premiers jours suivant le vêlage précédent. Le problème se ramène donc à un problème de régression logistique fonctionelle où on cherche à prédire une variable réponse dichotomique Y à partir de courbes X. En posant $\pi_i = P\left(Y = 1 | X = x_i(t); t \in T\right)$ la probabilité que la vache i soit gestante sachant l'évolution de sa production laitière en fonction du temps $x_i(t)$, on peut écrire le modèle suivant, pour $i = 1, ..., n$,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \int_0^T \beta(t) x_i(t) dt$$

avec pour hypothèse que les $\beta(t)$ et les courbes $x_i(t)$ sont dans le même espace de dimension finie [1]. On peut donc les écrire ainsi $\beta(t) = \sum_{q=1}^{p} b_q \Psi_q(t) = \mathbf{b}'\Psi$ et $x_i(t) = \sum_{q=1}^{p} c_{iq} \Psi_q(t) = \mathbf{c_i}'\Psi$ ce qui revient à un problème de régression fonctionnelle classique

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 \mathbf{I} + \mathbf{C\Phi b}$$

avec $\mathbf{C} = (c_{iq})$ et $\mathbf{\Phi} = \left(\Phi_{kq} = \int_T \Psi_k(t)\Psi_q(t)dt\right)$.

Différents packages R sont disponibles pour réaliser ce type d'analyse et nous présentons ici en détail les packages **fda** [2] et **fda.usc** [3]. Nous commençons par présenter le format des données requis pour l'utilisation de ces packages, puis nous détaillons l'utilisation de la fonction de régression logistique fonctionnelle *fregre.glm*. Le package est illustré sur les données des bovins et les résultats sont comparés à ceux obtenus par régression logistique classique (non fonctionnelle), avec sélection de variables, qui est implémentée dans la fonction *glm* du package **stats**. Les résultats de 50 validations croisées sont comparés selon 3 critères, le taux global de bien classés, la sensibilité et la spécificité.

Références

[1] Cardot, H., Faivre, R., Goulard, M., (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. Journal of Applied Statistics, Vol. 30, 1185-1199.

[2] Ramsay, J. O., Silverman, B. W. (2006), Functional Data Analysis, 2nd ed., Springer, New York.

[3] Febrero-Bande, M., Oviedo de la Fuente, M., (2012), Statistical Computing in Functional Data Analysis : The R Package fda.usc, Journal of Statistical Software, 51(4), 1-28.

Données tronquées sous R dans les modèles linéaires simples et à effets mixtes

D. Thiam[a] et **G. Nuel**[a,b]

[a]Labo de Maths Appliquées (MAP5, CNRS 8145)
Université Paris Descartes
djeneba.thiam@gmail.com

[b] Institut des Maths et Interactions (INSMI)
CNRS Paris
gregory.nuel@parisdescartes.fr

Mots clefs : Données tronquées, modèle Tobit, modèles mixte, algorithme EM.

Nous nous intéressons à une variable réponse tronquée avec la possibilité d'avoir plusieurs seuils. Dans le cadre des modèles linéaires simples, le modèle Tobit [1], très populaire en économie permet la prise en compte des troncatures hautes et basses. Dans le cadre des modèles à intercept aléatoires, autrement appelé données de panel dans le domaine économique, Tobit adapte la présence d'effets aléatoires en utilisant une approximation de l'intégrale sur les effets aléatoires par la méthode de quadrature de Gauss Hermite [2]. D'autres alternatives sont possibles, via des algorithmes itératifs de type EM [3] , Marquart [4], Newton Ralphson [5, 6]. Ces derniers permettent une prise en compte plus générale des effets aléatoires.

Sous R la gestion des données tronquées dans les modèles linéaires se fait à l'aide des librairies `censReg` [7], `AER` [8]. Dans le cadre des données de panel, la librairie `censReg` dispose d'une option permettant la prise en compte de données à intercept aléatoire. Ces librairies sont rapides, et fournissent une estimation rapide des paramètres du modèle. Cependant certaines fonctionnalités pourraient être améliorées. Un premier point est celui du calcul des résidus du modèle en présence des troncatures. Le second point est la gestion des effets aléatoires et troncatures simultanément dans le cadre des modèles linéaires à effets mixtes avec troncatures.

L'objectif de ce travail est de présenter une approche plus générale pour la prise en compte des données tronquées en présences d'effets aléatoires. En effet en combinant l'algorithme EM, les lois conditionnelles et les sorties R des packages `lmer` [9] ou `censReg` [7], nous arrivons à obtenir une estimation approchée des paramètres du modèle linéaire à effets mixtes en présence de troncatures doubles ou multiples.

Considérons le modèle suivant:

$$\mathtt{y_{ij}} = \beta \mathtt{x_{ij}} + \mathtt{z_i} + \varepsilon_{\mathtt{ij}}$$

- $\mathtt{y_{ij}}$: variable de réponse avec possibilité de troncatures.
- $\mathtt{y} = (\mathtt{y_T}, \mathtt{y_0})$ ou $\mathtt{T} = \{i, j \text{ tq } \mathtt{y_{ij}} \text{ tronqué}\}$
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ résidus; $\mathtt{z_i}$ iid $\sim \mathcal{N}(0, \eta^2)$: effet aléatoire
- $\mathtt{x_{ij}}$: covariable ($\in \mathbb{R}^n$)

Nous nous proposons de résoudre ce problème d'estimation par l'algorithme EM. En considérant l'ensemble $(\mathtt{y_T}, \mathtt{z})$ comme donnée non observée, le paramètre θ du modèle à l'itération courante

est obtenu de la manière suivante:

$$M(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{z}} \int_{\mathtt{y_T}} \mathbb{P}(\mathtt{z}, \mathtt{y_T}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}|\mathtt{z}, \mathtt{y_T}; \theta') \, \mathrm{d}\mathtt{z} \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

Afin de mettre à jour le paramètre courant en dehors de la procédure d'estimation, nous posons le problème différemment: en remarquant qu'à `z` (rep. $\mathtt{y_T}$) donné la vraisemblance $\mathbb{P}(\mathtt{y_0}, \mathtt{z}|\theta)$ (resp. $\mathbb{P}(\mathtt{y_0}, \mathtt{y_T}|\theta)$) est celle obtenue par `lmer` (resp. `tobit`). Ainsi deux techniques permettent d'obtenir le paramètre θ à l'itération courante :
1) EM combiné avec `lmer` du package `lme4`

$$M_{\text{lmer}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{y_T}} \mathbb{P}(\mathtt{y_T}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}, \mathtt{y_T}; \theta') \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

2) EM combiné avec `tobit` des packages `censReg` ou `AER`

$$M_{\text{tobit}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{z}} \mathbb{P}(\mathtt{z}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}, \mathtt{z}; \theta') \, \mathrm{d}\mathtt{z}}_{Q(\theta'|\theta)}$$

Nous comparons ces deux techniques entres elles en terme d'estimation des paramètres, des résidus et des effets aléatoires. Ces méthodes sont aussi comparées aux alternatives R existantes comme `censReg` pour données panel, ou une implémentation d'un algorithme EM stochastique associé à du Gibbs sampling pour l'échantillonnage sous les lois conditionnelles.

Références

[1] J. Tobin (1958). Estimation of relationship for limited dependent variables. *Econometrica*, **26**, 24-36.
[2] J. Pan; R. Thompson (2003). Gauss-Hermite Quadrature Approximation for Estimation in Generalised Linear Mixed Models, **18**, 57-78.
[3] A. P. Dempster; N. M. Laird; D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm *Journal of the Royal Statistical Society* , **39**, 1-38.
[4] D. Marquardt (1963). Methods of Conjugate Gradients for Solving Linear Systems. *SIAM Journal on Applied Mathematics*, **11**, 1-462.
[5] F. Cajori(1911). Historical note on the Newton Raphson method of approximation. *Am. Math. Monthly*, **18**, 19-33.
[6] H. W. Richmond (1944). On the Newton-Raphson method of approximation *Edinburgh Math. Notes*, **44**, 5-8.
[7] A. Henningsen (2010). Estimating Censored Regression Models in R using the censReg Package. `http://cran.r-project.org/web/packages/censReg/vignettes/censReg.pdf`.
[8] C. Kleiber; A Zeileis (2010). Applied Econometrics with AER package version 1.1-7. `http://CRAN.R-project.org/package=AER`.
[9] D. Bates; M. Maechler; B. Bolker (2010).Package lmer. `http://cran.r-project.org/web/packages/c`

Rankclust: An R package for clustering multivariate partial rankings

Q. Grimonprez[a], **J. Jacques**[a,b] and **C. Biernacki**[a,b]

[a]Modal team
Inria Lille-Nord Europe
Villeneuve d'Ascq, France
quentin.grimonprez@inria.fr

[b]Laboratoire Paul Painlevé
Université Lille 1
Villeneuve d'Ascq
julien.jacques@polytech-lille.fr, christophe.biernacki@math.univ-lille1.fr

Keywords : model-based clustering, multivariate ranking, partial ranking.

1 Introduction

Ranking data occur when a number of subjects are asked to rank a list of objects $\mathcal{O}_1, \ldots, \mathcal{O}_m$ according to their personal order of preference. The resulting ranking can be designed by its *ordering* representation $x = (x^1, \ldots, x^m) \in \mathcal{P}_m$ which signifies that Object $\mathcal{O}_{x^h}$ is the hth ($h = 1, \ldots, m$), where $\mathcal{P}_m$ is the set of the permutations of the first m integers. These data are of great interest in human activities involving preferences, attitudes or choices like Politics, Economics, Biology, Psychology, Marketing, *etc.* For instance, the voting system *single transferable vote* occurring in Ireland, Australia and New Zeeland, is based on preferential voting.

2 Mixture of multivariate ISR model

Starting from the assumption that a rank datum is the result of a sorting algorithm based on paired comparisons, and that the judge who ranks the objects uses the insertion sort because of its optimality properties, [1] state the following ISR model:

$$\mathrm{p}(x; \mu, \pi) = \frac{1}{m!} \sum_{y \in \mathcal{P}_m} \pi^{G(x,y,\mu)} (1-\pi)^{A(x,y)-G(x,y,\mu)}, \tag{1}$$

where $\mu \in \mathcal{P}_m$ is a *location parameter* and $\pi \in [\frac{1}{2}, 1]$ is a *scale parameter*. The numbers $G(x, y, \mu)$ and $A(x, y)$ are respectively the number of good paired comparisons and the total number of paired comparisons of objects during the sorting process (see [1] for more details). Recently, [2] propose a model-based clustering algorithm for multivariate rankings, i.e. when a datum is composed of several rankings, potentially partial (when some objects have not been ranked). For this, they extend the ISR model by assuming that, given a group k, the components of a multivariate ranking are independent:

$$\mathrm{p}(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \mathrm{p}(x^j; \mu_k^j, \pi_k^j), \tag{2}$$

where the model parameter $\boldsymbol{\theta} = (\pi_k^j, \mu_k^j, p_k)_{k=1,\ldots,K\ ,j=1,\ldots,p}$ are estimated by the mean of a SEM-Gibbs algorithm. The resulting algorithm is able to cluster ranking data sets with full and/or partial rankings, univariate or multivariate. To the best of our knowledge, this is the only clustering algorithm for ranking data with a so wide application scope.

3 The Rankclust package

This algorithm has been implemented in C++ and is available through the **Rankclust** package for **R**, available on the author webpage[1] and soon on the CRAN website[2].

The main function `rankclust()` performs cluster analysis for multivariate rankings and is able to take into account partial ranking.
This function has only one mandatory arguments: `data`, which is a matrix composed of the observed ranks in their ordering representation. The user can specify the number of clusters (1 by default) he wants to estimate or provide a list of clusters numbers. In that case, the user can choose either the BIC or ICL criterion to select the best number of clusters among his list. The outputs of `rankclust()` are of different nature:

- the estimation of the model parameters as well as the 'distances' between the final estimation and the current value at each iteration of the SEM-Gibbs algorithm. These distances can be used as indicators of the estimation variability.
- the estimated partition. Additionally, for each cluster, the probability and the entropy for all the cluster's members are given. This information helps the user in its interpretation of the clusters.
- for each partial ranking, an estimation of the missing positions.

4 Application

The use of the **Rankclust** package will be illustrated by the analysis of the European countries votes at the Eurovision song contest from 2007 to 2012.

References

[1] C. Biernacki and J. Jacques. A generative model for rank data based on sorting algorithm. *Comput. Statist. Data Anal.*, 58:162–176, 2013.

[2] J. Jacques and C. Biernacki. Model-based clustering for multivariate partial ranking data. Technical Report 8113, Inria Research Report, 2012.

[1] `http://labomath.univ-lille1.fr/~jacques/`
[2] `http://cran.r-project.org/`

SOMbrero : Cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités

Laura Bendhaïba[a], **Madalina Olteanu**[a] et **Nathalie Villa-Vialaneix**[a,b]

[a]SAMM, Université Paris 1
F-75634 Paris - France
laurabendhaiba@gmail.com
{madalina.olteanu,nathalie.villa}@univ-paris1.fr

[b]INRA, UR875, MIAT
F-31326 Castanet Tolosan - France

Mots clefs : cartes auto-organisatrices, dissimilarités, graphes, classification, visualisation

Dans de nombreuses situations réelles, les individus sont décrits par des jeux de données multiples qui ne sont pas nécessairement de simples tableaux numériques mais peuvent être des données complexes (graphes, variables qualitatives, texte...). Un cas typique est celui des graphes étiquetés dans lequel les individus (les sommets du graphe) sont décrits à la fois par leurs relations les uns aux autres mais aussi par des attributs de natures diverses. Dans [5, 2], nous avons proposé d'utiliser des cartes auto-organisatrices [1] pour combiner classification et visualisation en projetant les individus étudiés sur une grille de faible dimension. Notre approche permet de traiter des données non numériques par le biais de noyaux ou de dissimilarités, et est basée sur une version stochastique de l'apprentissage de cartes auto-organisées, comme décrit dans [4, 3]. Les différentes dissimilarités sont combinées et la combinaison est optimisée au cours de l'apprentissage de la carte.
Nous avons testé notre approche sur un jeu de données simulé : dans celui-ci, les observations sont décrites par un graphe séparé en deux groupes denses de sommets (figure 1, en haut à gauche), les sommets étant étiquetés par des valeurs numériques de $\mathbb{R}^2$ tirées selon deux Gaussiennes (figure 1) ainsi que par un facteur à deux niveaux. Seules les trois informations permettent de retrouver les 8 groupes de sommets, représentés par 8 couleurs différentes sur la figure 1. La combinaison des trois informations sous la forme de trois tableaux de dissimilarités (longueur du plus court chemin entre deux sommets pour le graphe, distance euclidienne pour les étiquettes numériques et distance de Dice pour les facteurs) permet de retrouver les huit groupes initiaux avec une bonne précision et de bien les organiser sur la carte (figure 1, en bas à droite). L'apprentissage adaptatif des distances donne un poids prépondérant à la dissimilarité basée sur la valeur du facteur qui est la seule valeur non bruitée (figure 1).
La méthodologie proposée est en voie d'implémentation dans un package R appelé **SOMbrero**. La version 0.1 du package, disponible depuis mars 2013 propose l'implémentation de l'algorithme de cartes auto-organisatrices pour des données numériques simples ainsi que diverses fonctionnalités permettant l'interprétation (fonctionnalité graphique pour visualiser les niveaux des diverses variables, les valeurs des prototypes de la carte...). Le package n'est pas encore disponible sur le CRAN mais peut être téléchargé à http://tuxette.nathalievilla.org/?p=1099&lang=en (sources et compilation windows).

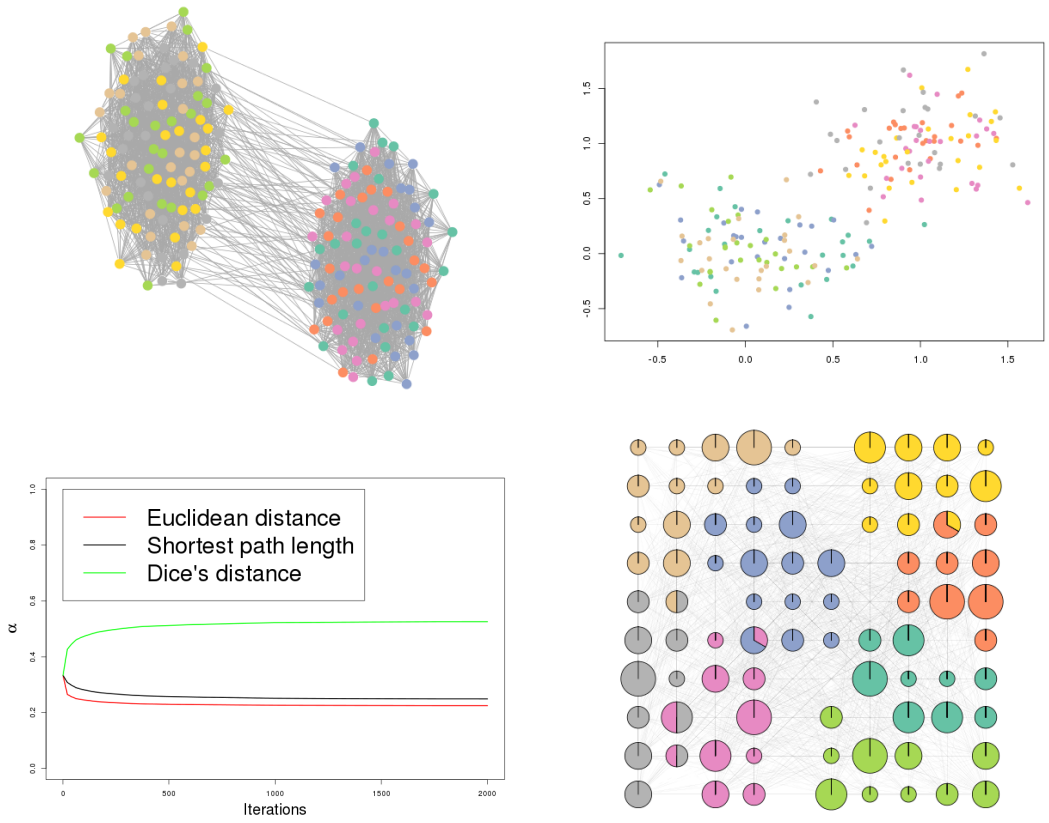


Figure 1: Données simulées : graphes et valeurs des étiquettes numériques des sommets (en haut à gauche et à droite). Évolution des poids des diverses dissimilarités (en bas à gauche). Carte finale obtenue (en bas à droite, les couleurs représentent les classes initiales, les aires des disques sont proportionnelles au nombre d'observations de la classe)

References

[1] T. Kohonen. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.

[2] M. Olteanu, N. Villa-Vialaneix, and C. Cierco-Ayrolles. Multiple kernel self-organizing maps. In M. Verleysen, editor, *XXIst European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, 2013. d-side publications. Forthcoming.

[3] M. Olteanu, N. Villa-Vialaneix, and M. Cottrell. On-line relational som for dissimilarity data. In P.A. Estevez, J. Principe, P. Zegers, and G. Barreto, editors, *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*, volume 198 of *AISC (Advances in Intelligent Systems and Computing)*, pages 13–22, Santiago, Chile, 2012. Springer Verlag, Berlin, Heidelberg.

[4] N. Villa and F. Rossi. A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph. In *6th International Workshop on Self-Organizing Maps (WSOM)*, Bielefeld, Germany, 2007. Neuroinformatics Group, Bielefeld University.

[5] N. Villa-Vialaneix, M. Olteanu, and C. Cierco-Ayrolles. Carte auto-organisatrice pour graphes étiquetés. In *Actes des Ateliers FGG (Fouille de Grands Graphes), colloque EGC (Extraction et Gestion de Connaissances)*, Toulouse, France, 2013.

The data.sample package: sampling as a big data mining tool

M. CORNEC [a] and J. DAS NIEVES [a]

[a] Data Operation Unit
CDISCOUNT
120-126 quai Bacalan – CS 11584
matthieu.cornec@cdiscount.com
jean.dasnieves@discount.com

Mots clefs: big data inference, sampling.

Not a single day without a newspaper article on the big data deluge. According to media coverage, hundreds of Teraoctets (To) would be waiting to be explored, and to deliver great value for consumers and companies. At cdiscount, French leading ecommerce website, we collect a dozen of To on a monthly basis.

At the same time, R, maybe the most popular statistical language is not ready for the Big Data Era. This native drawback is well known since R must load data sets into RAM.

Different strategies have been developed to tackle this challenge. Among them, we can quote: muscling in-house hardware, combining R and a big data relational data base language (such as Hive), the biglm package, the RSQLite package, ff package. Their main purpose is to run SQL like queries for data sets that do not fit into memory. Thus, they give accurate and deterministic results such as the sum or the mean of a variable.

In this poster, we defend the following strategy: as far as data analysis is concerned, sampling is a reliable, fast, and cheap data mining tool for big data. This statement can sound paradoxical since sampling is traditionally associated to the 20th century and to the theory of representative sampling. Nowadays, it is a common belief that we have access to exhaustive piece of information, so why sampling? The reason is the following: when it deals with modeling by opposition to reporting, the error induced is negligible in comparison with model errors, model noise, estimation error,

We introduce the data.sample package whose main function read.table.ds takes the location of big file as input and returns an object of class table.ds, which contains the sampled dataset together with the sampling weights. The main interest is that this strategy is not limited by the RAM size. By allowing the reuse of other R packages, it produces fast and actionable results.

To support our point view, we give theoretical insights following [1,2], simulation studies, and real data sets manipulations.

The data.sample package will be available on request from the authors.

Références

[1] A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan (2011). Bootstrapping big data. Big Learn, 2011.
[2] Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2011). A scalable bootstrap for massive data. *arXiv preprint arXiv:1112.5016*.
[3] T. J. Hastie, R.J. Tibshirani, and J.H. Friedman. The elements of statistical learning. Springer, 2009.

Génération de documents Word à partir de R : utilisation du package R2DOCX dans une plateforme statistique en milieu industriel.

David Gohel [a]

[a] Consultant R
Lysis Consultants
david.gohel@lysis-consultants.fr

Mots clefs : génération de document MS docx, R.

L'automatisation du reporting est un point d'optimisation essentiel dans le milieu industriel, à la fois pour diminuer le temps considérable consacré à l'édition de rapports mais également pour tendre vers une recherche reproductible et assurer la qualité et l'homogénéité des analyses. Il est de plus incontournable, pour assurer la communication des résultats au sein de l'entreprise, de présenter des analyses statistiques éditables, avec des mises en forme spécifiques.

Aujourd'hui plusieurs solutions de reporting éditable existent via l'utilisation de R. Cependant, elles ne sont pas pleinement satisfaisantes : difficulté d'intégration dans un contexte industrialisé, mises en forme parfois limitées, difficultés d'intégration des modèles de document (template Microsoft Word).

Afin de répondre au besoin d'un client, la société Lysis a développé le package R 'R2DOCX' qui permet la génération de document MS Word à partir de R dans un contexte industrialisé. Ce package s'appuie sur la bibliothèque open source java docx4j[1] qui offre la possibilité de créer et de manipuler des documents au format Microsoft Open XML[2]. Aucun autre composant logiciel n'est requis.
La première version de R2DOCX offre la possibilité d'automatiser la production de tableaux, de graphiques, de textes et de tables des matières. Le document final peut être personnalisé grâce à l'utilisation d'un modèle défini par l'utilisateur. Le document final est réalisé via l'ajout successif d'éléments à éditer, produits par R (tables, graphiques, textes) à partir de ce modèle et des styles prédéfinis.
La mise en forme et le contenu des tableaux sont facilement configurables par l'utilisateur. Il est possible de modifier les entêtes, de fusionner des cellules, de spécifier les formats des contenus (numériques, pourcentages, etc.) et des cellules (bordures, espace entre le contenu et les bordures, etc.).
L'insertion de paragraphes, qui permet d'ajouter du texte, permet également la mise en forme conditionnelle de sous-paragraphes. Finalement, il est également possible de remplacer des mots clefs dans le document modèle, comme par exemple le nom de l'auteur, le titre, la date etc.

L'ensemble de ces fonctionnalités sera illustré au travers d'exemples simples d'analyses descriptives.

Références

[1] docx4j. http://www.docx4java.org/
[2] Standard ECMA-376 Office Open XML File Formats. http://www.ecma-international.org/publications/standards/Ecma-376.htm

sexy-rgtk: a package for programming RGtk2 GUI in a user-friendly manner

Damien Leroux[a] and **Nathalie Villa-Vialaneix**[a,b]

[a] INRA, UR875, MIAT
F-31326 Castanet - France
damien.leroux@toulouse.inra.fr

[b] SAMM, Université Paris 1
F-75634 Paris - France
nathalie.villa@univ-paris1.fr

Keywords: Gtk2, RGtk2, GUI

There are many different ways to program Graphical User Interfaces (GUI) in R. [1] provides an overview of the available methods, describing ways to program R GUI with **RGtk2**, **qtbase** and **tcltk**. More recently, the package **shiny**, for building interactive web applications, was also released (the first version has been published on December, 2012).
The package **RGtk2** [2] is probably one of the most complete packages to program complex and highly customizable GUI. It is based on GTK2 (the GIMP Toolkit, http://www.gtk.org/), which is a multi-platform toolkit for creating Graphical User Interfaces. GTK2 offers a complete set of widgets and can be used to develop complete application suites working on Linux, Windows and Mac OS X. Although very flexible, each **RGtk2** interface results in a long script that has a counterintuitive syntax for most R users. For instance, the simple window of Figure 1[1] is obtained with the command lines provided in Figure 2 (left).

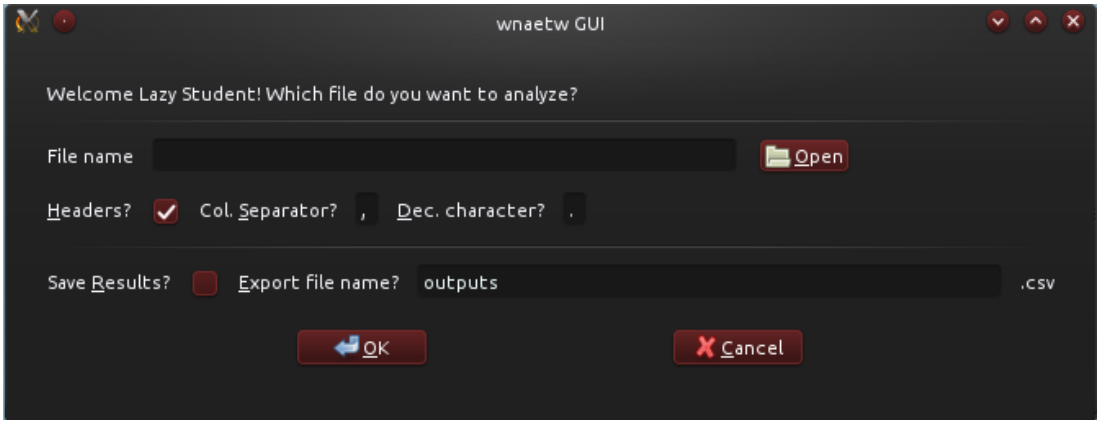


Figure 1: A simple GUI interface made with **RGtk2**.

One attempt to overcome the difficulty of the **RGtk2** syntax is the package **gWidgets** but, quoting its reference manual

> *"The excellent* **RGtk2** *package opens up the full power of the GTK2 toolkit, only a fraction of which is available though* **gWidgetsRGtk2**.*"*

[1] obtained on Kubuntu 12.04, Satanic Edition. The window's appearance differs depending on the OS and on the system's color configuration.

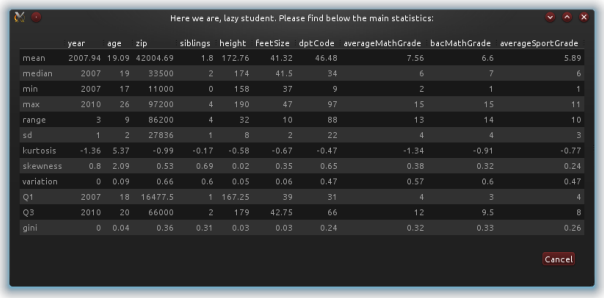
By automatically indexing all objects and methods available in **RGtk2**, we[2] developed a method for creating GTK2-based GUI, in a friendlier and more compact manner. Widgets are accessible with simple functions and options, as is more natural for a R language programmer. The window of Figure 1 is thus generated by the script provided in Figure 2 (right).

```
window <- gtkWindow()
window["title"] <- "wnaetw GUI"
vbox <- gtkVBoxNew(FALSE, 8)
vbox$setBorderWidth(24)
window$add(vbox)
hbox <- gtkHBoxNew(FALSE, 8)
vbox$packStart(hbox, FALSE, FALSE, 0)
label <- gtkLabelNew("Welcome Lazy Student!
  Which file do you want to analyze?")
hbox$packStart(label,FALSE,FALSE,0)
vbox$packStart(gtkHSeparatorNew(), FALSE, FALSE, 0)
hbox <- gtkHBoxNew(FALSE,8)
vbox$packStart(hbox, FALSE, FALSE, 0)
label <- gtkLabelNewWithMnemonic("_File name")
hbox$packStart(label,FALSE,FALSE,0)
filename <- gtkEntryNew()
filename$setWidthChars(50)
label$setMnemonicWidget(filename)
hbox$packStart(filename,FALSE,FALSE,0)
buttonOpen <- gtkButtonNewFromStock("gtk-open")
gSignalConnect(buttonOpen, "clicked", openFile)
hbox$packStart(buttonOpen,FALSE,FALSE,0)
# ... (script is cut at 1/3 of its length)
```

```
main <- Window(title="wnaetw GUI",
  contents=Rows(
    Label("Welcome Lazy Student! Which file do you want
        to analyze?"),
    br,HSeparator(), br,
    LabeledWidget("_File location ",
      Entry(use.name='filename',width.chars=50),
      mnemonic=T),
    Button(from.stock="gtk-open", on('clicked',run=
      chooseFile()), fill=F),br,
  # ... (script is cut at 1/3 of its length))
```

Figure 2: **RGtk2** (left) vs **sexy-rgtk2** (right).

This method has been used for recoding in a very short and simple manner the basic GUI of the (toy) package **wnaetw** [3]. Also a function has been developed to ease the use of the function `rGtkDataFrame`. A `data.frame` object `res` can thus be displayed in a window with the single command `DataFrame(res)` instead of having to define individually each column renderer. This feature is illustrated in Figure 3. The method should be released as a package next summer but the first scripts, without documentation, as well as the demo code, are available upon request.




```
performWmtw(main$filename$text, main$headers$active,
  main$sep$text,main$dec$text, main$quote$active,
  main$saveres$active,main$savename$text)
#...
performWmtw <- function(fn, headers, sep, dec, quote,
      save, sn) {
  # ... reading file
  res <- applyWmtw(my.data)
  ## GUI part
  resGUI <- Dialog(title="Here we are, lazy student.
    Please find below the main statistics:")
  DialogRows(resGUI,DataFrame(res, row.names=TRUE),br,
    pack(Button(from.stock='Cancel', on('clicked',run=
    resGUI$destroy())),expand=F, fill=T, padding=20,
    whence="end"))
```

Figure 3: Use case example for the function `DataFrame`.

References

[1] M. Lawrence and J. Verzani. *Programming Graphical User Interfaces in R*. CRC The R Series. Chapman & Hall, June 2012.

[2] L. Michael and T.L. Duncan. RGtk2: A graphical user interface toolkit for R. *Journal of Statistical Software*, 37(8):1–52, 2010.

[2] The authors did not contribute equally to the work: Damien developped the method to extract and interface **RGtk2** objects and methods, whereas Nathalie was the friendly useR and beta testeR.

[3] "What Nicolas's Teacher Wants", described at http://tuxette.nathalievilla.org/?p=885&lang=en.

R-dyndoc une alternative à Sweave

R. Drouilhet

Laboratoire Jean Kuntzmann, Grenoble
remy.drouilhet@upmf-grenoble.fr

Mots clefs : Statistique, Rapport automatique.

1 Motivation

Le projet R-dyndoc a commencé il y a une quinzaine d'années avec pour premiers objectifs :

- **Enseignement :** gain de temps lors de la rédaction des sujets d'examens et préparation de documents de cours.
- **Consulting statistique :** idée de proposer des outils de génération automatique de rapports en intégrant les résultats extraits de traitements R.
- **Programmation :** création d'un outil permettant la combinaison de plusieurs langages aux caractéristiques complémentaires (par exemple, la facilité du R à la manipulation et le traitement de données et la manipulation aisée en ruby des chaînes de caractères).

Grâce à la facilité à "embarquer" le système R dans des langages de programmation basés sur des API C (mon préféré étant ruby), il est aujourd'hui relativement facile de développer dans l'un de ces langages une alternative à Sweave (intégralement développé en R). L'un des avantages est certainement la possibilité offerte d'utiliser un plus grand nombre de librairies.
Maintenant, R-dyndoc est un outil développé en ruby avec pour principales caractéristiques :

- **système de modèles ("templating system") :** basé sur un langage de balises (de type "domino") suffisamment exotique pour s'intègrer facilement dans tout document au format "human readable" (tel que latex, html, ...) afin d'en dynamiser le contenu.
- **langage (de script) :** permettant d'intégrer pleinement les langages ruby et R (avec possibilité d'intégration de python et tout autre langage interfaçable en C).

Les principaux avantages de R-dyndoc sont ses possibilités :

- de créer (dans un *mode développeur*) des librairies regroupant des collections de fonctions et objets R-dyndoc (spécifiquement dédiés à la génération de parties textuelles relatives à certaines expertises) facilement utilisables (dans un *mode utilisateur*) à la rédaction finale de rapports.
- de rassembler dans un même document R-dyndoc (plus communément appelé "template" R-dyndoc) à la fois des traitements (R et ruby) effectués dans une analyse statistique (par exemple) et le contenu du rapport à diffuser au format pdf ou html.
- de créer, à partir d'un unique "template" R-dyndoc, plusieurs documents (éventuellement dans différents formats latex et html).

Notez que le R-dyndoc n'est pas encore disponible en version stable. Il est toutefois possible de le tester en installant deux bibliothèques ruby (appelés gems) ainsi que deux packages R (voir page web http://dyndoc.upmf-grenoble.fr/DyndocInstall.html).

2 Quelques exemples

Pour conclure, illustrons à travers quelques exemples basiques (malheureusement non commentés ici par manque de place) quelques fonctionnalités offertes par R-dyndoc.

```
[#r<] mister <- "misteR" #R code
[#rb<] mister = "mister" #ruby code
[#<] This is not output!
   {#def]hello[#,]name[Miss][#>]
      [Hello #{name}]
   [#}
[#>]
      [from Dyn, {#hello#}
      |from Dyn, {#hello]Mister[#}
      |from ruby, hello :{mister}
      |from R, hello :r{mister}]
```

Résultat :

```
from Dyn, Hello Miss
from Dyn, Hello Mister
from ruby, hello mister
from R, hello misteR
```

```
[#=] docs [part2,part3,part1]
[#>]{#case]#{docs}
   [#when]part1[#>]Partie~ I
   [#when]part2[#>][Partie~ II<\n>]
   [#when]part3[#>]Partie~ III
   [#}|
```

Résultat :

```
Partie~II
Partie~III
Partie~I
```

```
[#>]for loop in R:
[#R>] for(cpt in 1:4) {
   {#>]item:r{cpt} [#>}
}
[#>]<\n>sapply loop in R:
[#R>] sapply(5:8,function(cpt)
   {#>]item#r{cpt} [#>}
)
```

Résultat :

```
for loop in R:
item1 item2 item3 item4
sapply loop in R:
item5 item6 item7 item8
```

```
[#=]toto[TOTO]
[#=]toto$[c(1,3,2)]
[#>]Before: #{toto}  AND #{toto$}
[#R<]
<toto:> = tolower(<toto:>)
<toto$>[1] = 4
[#>]After: #{toto}  AND #{toto$}
```

Résultat :

```
Before: TOTO  AND 1.0,3.0,2.0
After: toto  AND 4.0,3.0,2.0
```

Les sorties suivantes montrent le même exemple en Sweave puis en R-dyndoc :

```
\documentclass{article}
\begin{document}
«»=
a<-c(1,3,2)
a
@
La moyenne est \Sexpr{mean(a)}.
\end{document}
```

```
{#rverb]
a<-c(1,3,2)
a
[#}
La moyenne est :r{mean(a)}.
```

Index des auteurs

Aissani, Djamil 22
Anton, Halina 85
Bahram, Seiamak 26
Bendhaïba, Laura 99
Bertrand, Frédéric 26
Bicout, Dominique 8
Biernacki, Christophe 97
Bonadona, Valérie 73
Bonnefoy, Cyril 35
Boumaza, Rachid 22
Briatte, François 84
Caeiro, Frederico 24
Cardot, Hervé 93
Celeux, Gilles 64
Charles, Sandrine 80
Charpentier, Arthur 34
Chasset, Pierre-Olivier 71
Chauveau, Didier 40
Cheylus, Anne 31
Chine, Karim 36
Chuffart, Florent 32
Clavel, Julien 10
Collin, Jean-François 102
Cornec, Matthieu 101
Das Nieves, Jean 101
De Gaudemaris, Régis 8
Deguen, Séverine 69
Delbecq, Francoise 12
Delespierre, Tiba 78
Delignette-Muller, Marie Laure 67, 80
Dray, Stéphane 6, 20, 53, 58
Drouet, Youenn 73
Drouilhet, Rémy 105
Dufour, Anne-Béatrice 20, 53, 58
Dutang, Christophe 67
Escarguel, Gilles 10
Falissard, Bruno 16, 78
Fargier, Raphaël 31

Francois, Romain 42
Frindel, Carole 29
Gallic, Ewen 34
Gallopin, Mélina 64
Gayet, Anne 86
Genolini, Christophe 16
Giraud, Timothée 4
Godet, Julien 85
Gohel, David 102
Gombin, Joël 66
Grimonprez, Quentin 97
Husson, François 2
Irisson, Jean-Olivier 76
Jacques, Julien 97
Jaffrézic, Florence 60, 64
Julien Laferrière, Alice 20
Julien-Laferriere, Alice 6
Jung, Nicolas 26
Khaneboubi, Mehdi 56
Kihal, Wahida 69
Knoblauch, Kenneth 14
Kon Kam King, Guillaume 80
Labenne, Amaury 18
Lalloué, Benoit 69
Lasset, Christine 73
Laurent, Alexandre 43
Le Meur, Nolwenn 69
Le Pape, Gilles 88
Leroux, Damien 103
Lobry, Jean 53, 58
Longis, Sandrine 88
Maalouf, Jean-Paul 88
Madelin, Malika 35
Marot, Guillemette 60
Mauguen, Audrey 43
Mazo, Gildas 33
Mazroui, Yassin 43
Mely, Yves 85

Merceron, Gildas 10
Michel, Carine 12
Monnez, Jean-Marie 69
Montaudouin, Séverine 88
Mousset, Sylvain 53, 55
Myriam, Maumy-Bertrand 26
Nazir, Tatjana 31
Nedellec, Raphaël 75
Noel, Yvonnick 51
Nuel, Gregory 95
Olteanu, Madalina 99
Padilla, Cindy 69
Penel, Simon 53
Petev, Ivaylo 84
Pham, Van Trung 33
Philipps, Viviane 91
Piedvache, Céline 78
Pierre-Jean, Morgane 62
Pingault, Jean-Baptiste 16
Proust-Lima, Cécile 91
Rau, Andrea 60, 64
Rieutort, Delphine 8
Riou, Jeremie 47
Ritschard, Gilbert 45, 89
Rondeau, Virginie 43
Rouanet, Anaïs 29
Rousseau, David 29
Rousseaux, Emmanuel 45
Saadi, Habib 38
Sauder, Cécile 93
Sautet, Philippe 12
Siberchicot, Aurélie 6, 20
Strupler, Nehemie 82
Sueur, Jérôme 49
Thiam, Djeneba 95
Thioulouse, Jean 6, 20, 58
Thurin, Jean-Michel 78
Thurin, Monique 78

Vallat, Laurent 26
Vandekerkhove, Pierre 40
Veber, Philippe 80
Vieuille, Caroline 88
Villa-Vialaneix, Nathalie 99, 103
Wickham, Hadley 1
Yousfi, Smail 22
Zaffran, Jeremie 12
Zmirou-Navier, Denis 69