

Livret des résumés

6^{èmes} RENCONTRES R



Anglet - 28/30 juin 2017

UNIVERSITÉ DE PAU ET
DES PAYS DE L'ADOUR

Campus de Montaury

ANGLET

<http://angletr2017.com>



Table des matières

R Package UP : Distribution Universelle de Prédition pour les métamodèles, Ben Salem Malek	1
Réduction de dimension en apprentissage non supervisé., Baron Alexis [et al.]	4
Improvements in GUTS model implementation under a Bayesian framework with R and JAGS, Baudrot Virgile [et al.]	6
Investigating Gene- and Pathway-environment Interaction analysis approaches, Broc Camilo [et al.]	8
GADAG : un paquet R dédié à l'inférence de Graphes Acycliques Dirigés par maximum de vraisemblance pénalisé, Champion Magali [et al.]	10
Evaluation du package Heemod et mise en place d'un tableau de synthèse en Shiny pour l'analyse de coût-efficacité., Carranza Alarcon Yonatan [et al.]	12
Approche bloc en ACP group-sparse: le package sparsePCA, Chavent Marie [et al.]	16
productivity : un package dédié aux calculs d'indices de productivité par la méthode DEA, Dakpo K Hervé [et al.]	17
Identification des bassin d'habitat des Pyrénées-Atlantiques, Damitio Emma	19
Analyse des débats sur le secret bancaire au Parlement suisse avec R, Deville Marion [et al.]	21

Un outil de gestion de scénarios d'étude statistique en R, Dhorne Thierry [et al.]	23
Analyse spatiale multi-échelles de données écologiques avec adespatial, Dray Stéphane	25
La couleur dans l'environnement R : un survol, Eterradoissi Olivier	26
Efficient simulation of complex queueing systems with the R package queue-computer, Ebert Anthony [et al.]	27
Les API, un enfeR ?, Fay Colin	29
1 dataset, 10 dataviz, Fay Colin	30
Interface Homme Machine : Data Management, Genolini Christophe	31
Une interface R-Shiny pour le choix de typologies sur variables qualitatives, Gouss-eff Matthieu	33
Encapsuler une application R avec Docker, Guyader Vincent	35
SI-APIMODEL, un système d'information scientifique conçu avec R, Shiny, RPostgreSQL/PostGIS, dplyr, DT, leafletR, ..., Haddad Abdelmalek	36
The #SurfeR project : visualiser twitter avec R, Holtz Yan	38
Clouds, containers and R, towards a global hub for reproducible and collaborative open data science, Karim Chine	41
Etude des règles d'assemblages des communautés : présentation d'une nouvelle approche méthodologique, Legras Gaëlle [et al.]	43
naniar: Data structures and functions for consistent exploration of missing data, Nicholas Tierney [et al.]	45

Projet ART, Paroissin Christian [et al.]	46
C3CO - Un package pour l'inference de la clonalite des cellules cancéreuses à partir du nombre de copies d'ADN, Pierre-Jean Morgane	50
The R package bigstatsr: Memory- and Computation-Efficient Statistical Tools for Big Matrices, Privé Florian	52
Visualisation interactive d'arbres de décision avec visNetwork, Robert Titouan [et al.]	53
Outil d'interprétation de score, Salette Elena	54
Comment je suis devenue crolute, Salmon Maëlle [et al.]	56
Linkspotter : outil interactif d'exploration et de visualisation de corrélations, Samba Alassane	58
Fouille des données du jeu Mathador, Saumard Matthieu	59
Ibbe-shiny : un exemple de serveur shiny fait-maison au service de la recherche locale, Siberchicot Aurélie [et al.]	61
Marey-Map Online : une application shiny pour estimer les taux de recombinaison grâce à l'approche par carte de Marey., Siberchicot Aurélie [et al.]	63
Travaux sous tension : traitement d'un volume important de données et outils de visualisation, Thieurmel Benoit [et al.]	65
FactoInvestigate - Description automatique de résultats d'analyse factorielle, Thuleau Simon [et al.]	67
Modélisation bayésienne d'une chronologie d'évènements archéologiques et analyse des chaines de Markov à l'aide du package 'ArchaeoPhases', Vibet Marie-Anne [et al.]	69
Liste des auteurs	71

R Package UP : Distribution Universelle de Prédiction pour les méta-modèles

M. BEN SALEM^{a b}

^a Institut Fayol

Mines de Saint-Étienne, UMR CNRS 6158, LIMOS

158 Cours Fauriel, 42023 Saint-Étienne Cedex 2

malek.ben-salem@emse.fr

^b DesignXplorer Team

ANSYS, Inc

11 avenue Albert Einstein Villeurbanne 69100

Mots clefs : Prédiction, Méta-modèle, Planification séquentielle

Les méta-modèles (ou surfaces de réponse) sont devenus des outils répandus en ingénierie et en recherche. Ils remplacent des fonctions coûteuses s (bâtie à partir d'une expérience ou d'un modèle numérique). Ces modèles sont construits en se basant sur un jeu de données de n observations $z_j = (x_j, y_j)$ où $1 \leq j \leq n$ et $y_j = s(x_j)$. L'objectif des méta-modèles est d'estimer une fonctionnalité de s (minimum, ligne de niveau, modèle de remplacement) en utilisant un méta-modèle \hat{s} . L'utilisateur cherche un meilleur compromis entre la précision de l'estimation et le nombre d'appels à la fonction s . Par conséquent, l'échantillonnage du jeu de données, $(x_j)_{1 \leq j \leq n}$, est crucial. Nous nous intéressons aux stratégies séquentielles d'échantillonnage. Ces méthodes découlent généralement d'une estimation de l'incertitude liée aux prédictions des méta-modèles probabilistes. Ici, nous présentons le *package R* "UP" qui fournit une distribution universelle de prédiction (UP) [1]. Cette distribution associe une incertitude de prédiction pour tout méta-modèle. Elle est basée sur les prédictions des sous méta-modèles construits par une méthode de ré-échantillonnage telle que la validation croisée. Par exemple, la définition de la distribution universelle ci-dessous est basée sur la méthode "*Leave-One-Out Cross-Validation*" (LOO-CV).

Définition 1 La distribution universelle de prédiction est la distribution empirique pondérée suivante :

$$\mu_{(n,\mathbf{x})}(dy) = \sum_{i=1}^n w_{i,n}(\mathbf{x}) \delta_{\hat{s}_{n,-i}(\mathbf{x})}(dy). \quad (1)$$

où $w_{i,n}(\mathbf{x})$ est un poids basé sur les distances entre \mathbf{x} et $(\mathbf{x}_j)_{1 \leq j \leq n}$ et $\hat{s}_{n,-i}(\mathbf{x})$ est la prédiction du sous méta-modèle construit sur le jeu de données en retirant le i^{eme} point.

La distribution UP peut être utilisée pour tous les méta-modèles : probabilistes ou déterministes. Le *package* "UP" permet l'utilisation directe du krigeage et de machines à vecteurs supports (SVM) dans le contexte de la prédiction universelle. Il est possible également d'enrichir la liste des méta-modèles en implémentant un adaptateur qui contient principalement deux méthodes : (`train` pour l'apprentissage et `predict` pour la prédiction).

Pour les méthodes de ré-échantillonnage, il permet l'utilisation directe du LOO-CV et de la validation croisée k -fold-CV. De surcroît, le *package* "UP" offre la possibilité de spécifier une autre méthode de ré-échantillonnage. En effet, l'utilisateur peut fournir ses sous échantillons.

La liste des appels pour avoir la distribution de prédiction est simple et directe. Il suffit de choisir une méthode de ré-échantillonnage et un méta-modèle et de construire une classe dédiée à la distribution UP (Voir Figure 1).

```

> krig      <- krigingsm$new()
> resampling <- UPClass$new(x,y,Scale =T)
> upsm       <- UPSM$new(sm= krig, UP= resampling )
> prediction <- upsm$upredict(xverif)
> plot1D(xverif, prediction,x, y)

```

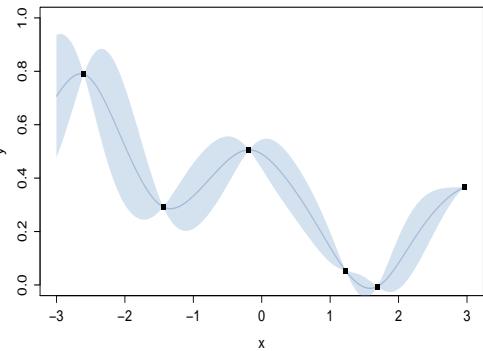


FIGURE 1 – Gauche : lignes de commandes. Droite : Figure correspondante, représentant : les points du plan d’expériences (carré noir), la prédition (ligne foncée) et la région délimitée par $\hat{s}_n(\mathbf{x}) \pm 3\hat{\sigma}_n(\mathbf{x})$ (bleu transparent) où $\hat{\sigma}_n(\mathbf{x})$ est la UP variance.

La prédition universelle permet d’étendre l’utilisation des stratégies séquentielles à des méta-modèles non probabilistes. Parmi ces stratégies, le *package “UP”* fournit une implémentations de (UP-EGO) “*Universal Prediction Efficient Global Optimization*” qui est l’extension de l’algorithme EGO [2] au cadre de la prédition universelle. On propose également un algorithme de raffinement automatique “*Universal Prediction Surrogate Modeling Adaptive Refinement Technique*” (UP-SMART). Des tests ont montré que ces algorithmes sont efficaces pour différent problèmes et permettent une utilisation plus large pour les méta-modèles.

Le package est flexible et peut supporter d’autres types d’échantillonnages et de méta-modèles. Il s’adapte aussi bien aux classes R6 qu’aux appels classiques. Des fonctionnalités de post-traitement sont également disponibles.

Références

- [1] Ben Salem, M., Roustant, O., Gamboa, F., and Tomaso, L. (2015). Universal Prediction Distribution for Surrogate Models. *arXiv preprint arXiv :1512.07560*.
- [2] Jones, D.R., Schonlau, M., and Welch, W.J. (1998) Efficient global optimization of expensive black-box functions, *Journal of Global Optimization*, **13** , 455-492
- [3] Roustant, O., Ginsbourger, D. and Deville, Y., (2012). DiceKriging, DiceOptim : Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1), p.54p.

Réduction de dimension en apprentissage non supervisé

Alex Mourer^a and Alexis Baron^a

^a M2 Techniques d'Information et de Décision dans l'Entreprise
SAMM, EA 4543, Université Paris 1 Panthéon Sorbonne

90 rue de Tolbiac, 75013, Paris

moureralex@gmail.com, alexis.b.baron@gmail.com

Mots clefs : apprentissage non supervisé, réduction de dimension, clustering.

Notre étude porte sur la réduction de dimension, et plus particulièrement sur l'évaluation de l'importance et la sélection de variables discriminantes, dans un contexte d'apprentissage non supervisé. Pour ce faire, nous utilisons un jeu de données [1] tiré d'une enquête sociologique réalisée auprès de 1010 jeunes slovaques âgés de 18 à 30 ans. Nous avons à disposition 150 variables qui correspondent à 150 questions sur les peurs, les habitudes, ou encore les goûts de ces jeunes. Chaque observation correspond à une réponse, et chaque réponse est donnée par une note comprise entre 1 et 5. Nous cherchons donc à sélectionner les variables les plus discriminantes, c'est à dire celles qui permettent de regrouper au mieux les individus similaires, et à réduire au maximum la dimension du tableau de données, en essayant de perdre un minimum d'information.

La méthodologie mise en place peut être décrite par les étapes ci-dessous.

1. Utilisation d'heuristiques et de méthodes stepwise

Un premier clustering \mathcal{C}_0 est réalisé à partir de l'ensemble des données, il servira de référence pour mesurer les éventuelles pertes d'information induites par la réduction de dimension. De plus, le nombre de cluster est dorénavant fixe pour le reste de la partie. Nous faisons ici l'hypothèse que le meilleur clustering est issu de l'ensemble des variables à disposition.

Par la suite, deux méthodes sont utilisées pour quantifier l'importance d'une variable :

- Les variables sont supprimées à tour de rôle, une à la fois, et le nouveau clustering est comparé à la référence \mathcal{C}_0 ;
- Pour chacune des variables, les valeurs sont permutées aléatoirement, et le clustering obtenu est alors comparé à celui issu des données sans permutation.

Les indicateurs utilisés pour mesurer l'impact des variables sur le clustering sont le pourcentage de variance expliquée, calculé comme le rapport entre la dispersion inter-classes et la dispersion totale, ainsi qu'une mesure de précision qui informe sur le pourcentage d'individus ayant changé de classe entre les deux clusterings à comparer.

Pour réaliser le clustering, deux algorithmes sont sélectionnés, le k -means [2] et les cartes auto-organisées [3], ces choix ayant été faits respectivement par rapport à la rapidité de l'un et aux bonnes propriétés de visualisation de l'autre. Par ailleurs, l'utilisation de deux algorithmes différents permet d'assurer plus de fiabilité aux résultats, indépendamment de la méthode de clustering employée.

A partir des méthodes évoquées précédemment, un groupe de variables jugées moins influentes individuellement peut être isolé. Afin de les retirer définitivement des variables influentes pour le clustering, nous utilisons une méthode de type stepwise. Les résultats obtenus peuvent être visualisés, par exemple, à l'aide d'une ACP.

2. Utilisation de variables latentes

La deuxième méthodologie mise en place pour essayer de réduire la dimension du jeu de données, une fois les variables jugées peu influentes sur le clustering déjà écartées, fait appel à une technique introduite dans [4]. Cette méthode se base sur la création de variables latentes. Elle permet de regrouper des variables fortement corrélées entre elles, ayant également l'avantage de tenir compte du signe de la corrélation. Cette dernière propriété peut-être utilisé pour interdire le regroupement de variables ayant des corrélations opposées et de ce fait cela évite de créer des concepts difficilement interprétable.

La spécificité de cette méthode est de s'apparenter à la fois aux algorithmes de clustering, en produisant une partition de l'ensemble des variables, et aux méthodes factorielles d'analyse de données. Une fois appliquée sur les 120 variables pré-sélectionnées de nos données, elle permet de créer une soixantaine de concepts latents. Ces concepts ont un sens, et la perte d'information induite par ce regroupement est contrôlée.

3. Utilisation d'arbres et forêts aléatoires non-supervisés

Enfin, dans la dernière partie de notre étude, des versions non-supervisées d'arbres et de forêts aléatoires sont utilisées pour sélectionner les variables les plus discriminantes. Dans un premier temps, l'algorithme Divclust [5] est appris sur les données. Son principe est similaire à celui des arbres de classification de type CART [6], chaque noeud étant défini dans ce cas par la variable qui maximise la variance inter-classes. Dans un second temps, les arbres Divlclust sont généralisés selon les mêmes idées que les forêts aléatoires, avec des tirages aléatoires parmi les variables et du bootstrap parmi les observations.

Références

- [1] <https://www.kaggle.com/miroslavabo/young-people-survey>
- [2] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* 28, 100–108.
- [3] Madalina Olteanu, Nathalie Villa-Vialaneix. On-line relational and multiple relational SOM. *Neurocomputing*, Elsevier, 2015, 147, pp.15-30.
- [4] Vigneau, E. and Qannari, E. M. (2003). Clustering of variables around latent components. *Comm. Stat. - Simul Comput* , 32(4), 1131 :1150
- [5] Marie Chavent, Yves Lechevallier, Olivier Briant. DIVCLUS-T : a monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis*, Elsevier, 2007, 52 (2).
- [6] L. Breiman, J.H. Friedman, R.A. Olshen, , and C.J Stone. Classification and Regression Trees. *Wadsworth, Belmont, Ca*, 1983.

Packages utilisés

- SOMbrero, dplyr, randomForest, ggplot2, doMC, doParallel, FactoMineR factoextra, Clust-VarLV

Improvements in GUTS model implementation under a Bayesian framework with R and JAGS

V. Baudrot^a and S. Charles^a

^aUniv Lyon, Université Lyon 1, UMR CNRS 5558, Laboratoire de Biométrie et Biologie Évolutive
UCB Lyon 1 - Bât. Grégor Mendel - F-69100 Villeurbanne, France
virgile.baudrot@univ-lyon1.fr
sandrine.charles@univ-lyon1.fr

Keywords: TK-TD models, survival models, Environmental Risk Assessment, package R 'morse'.

The application of toxicokinetic-toxicodynamic (TK-TD) modeling proved to be of particular interest in strengthening the Environmental Risk Assessment (ERA) of chemicals compounds (e.g., REACH dossier accounting for toxicity of industrial discharge, evaluation of impacts of Plant Protection Products (PPPs),...). TK-TD models simulate the time-course of processes leading to toxicity at the level of organisms. They include all mechanisms from external concentration to internal kinetics of compounds (e.g., exposure, uptake, elimination, biotransformation, internal distribution) named the toxicokinetic part, and translate the internal concentration into alteration of cells and organs functioning that can eventually lead to a toxic effect at the organism level (e.g., mortality, reduced reproduction, abnormal behavior) then affecting population dynamic [1,2].

For survival analysis of organisms in response to a chemical stressor, the General Unified Threshold model of Survival (GUTS) is today recognized as a suitable and powerful TK-TD framework incorporating two complimentary death mechanisms: Stochastic Death (GUTS-SD) and Individual Tolerance (GUTS-IT), from which a large range of existing models can be derived [1,2]. Intergovernmental institutions as the OECD have acknowledge the necessity of TK-TD models for ERA improvement [3], but while an integrative modelling as GUTS offers an efficient theoretical approach, its practical use is challenging (from model implementation to parameters estimation).

In order to ease the use of standard models associated with robust statistical methods in ERA, the web-platform MOSAIC (<http://pbil.univ-lyon1.fr/software/mosaic/>) was developed based on the 'morse' R package (<https://cran.r-project.org/web/packages/morse/index.html>). Both MOSAIC and 'morse' allow for visualizing and analyzing data from standard toxicity tests. Most of the analyses are provided under a Bayesian framework with JAGS. To date, concerning TK-TD models, only GUTS-SD is available in 'morse' for survival data collected through time under constant exposure [4]. In order to improve the 'morse' functionalities (and further the MOSAIC offer), we first implemented the 'GUTS-IT' under constant exposure, and then enhance both GUTS-SD/IT algorithms to allow parameter inference under variable time-course exposure concentration profiles [5]. Compared to other implementations of GUTS models [5], the 'morse' package allows a fully Bayesian approach for any GUTS model.

References

- [1] Ducrot, V., Ashauer, R., Bednarska, A. J., Hinarejos, S., Thorbek, P. and Weyman, G. (2016) Using toxicokinetic-toxicodynamic modeling as an acute risk assessment refinement approach in vertebrate ecological risk assessment. *Integrated environmental assessment and management*. **12**, 32–45
- [2] Jager, T., Albert, C., Preuss, T. G. and Ashauer, R. (2011). General Unified Threshold Model of Survival - a Toxicokinetic-Toxicodynamic Framework for Ecotoxicology. *Environmental Science & Technology, ACS Publications*, **45**(7), 2529-2540

-
- [3] Organisation for Economic Co-operation and Development (OECD) (2006) Environment health and safety publications. *Series on testing and Environmental Science & Technology assessment No. 54*. Current approaches in the statistical analysis of ecotoxicological data: a guidance to application. 1–147.
- [4] Delignette-Muller, M. L., Ruiz, P. and Veber, P. (2017) Robust fit of toxicokinetic-toxicodynamic models using prior knowledge contained in the design of survival toxicity tests. *Environmental Science & Technology, ACS Publications*.
- [5] Albert, C., Vogel, S. and Ashauer, R. (2016) Computationally Efficient Implementation of a Novel Algorithm for the General Unified Threshold Model of Survival (GUTS) *PLoS Comput Biol, Public Library of Science*, **12**, e1004978.

Investigating Gene- and Pathway-environment Interaction analysis approaches

C. Broc^a , M. Evangelou^b , T. Truong^c and B. Liquet^d

^aLaboratoire de Mathématiques et de leurs Applications de Pau
University of Pau et des Pays de l'Adour
Anglet, FRANCE
camilo.broc@univ-pau.fr

^bDepartment of Mathematics, Faculty of Natural Sciences

^bDepartment of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine
Imperial College
London, UK
m.evangelou@imperial.ac.uk

^cCancer and Environment team, CESP (Center for Research in Epidemiology and Population Health), INSERM
University Paris-Saclay, University Paris-Sud
Villejuif, FRANCE
therese.truong@inserm.fr

^dLaboratoire de Mathématiques et de leurs Applications de Pau
University of Pau et des Pays de l'Adour
Pau FRANCE
benoit.liquet@univ-pau.fr

Keywords : Gene-environment interactions, Generalized Linear Models, Pathway analysis, Resampling methods

Several approached were developed in order to complete the agnostic genome-wide association studies (GWAS) in the discovery of additional genetic risk factors or to provide additional insight into the mechanisms involved in the studied disease. One such approach is the gene-set analysis (GSA) also referred as pathway analysis, that consists in aggregating signals from variants to genes, which are in turn aggregated to sets of genes involved in a particular biological process. Although a number of statistical approaches have been proposed in the literature GSA, not a lot of work has been conducted for investigating $G \times E$ interaction at the gene and pathway- levels. We aim at extending existing GSA methods to the analysis of gene-environment interaction.

This paper explores gene- and pathway-environment interaction analysis by comparing Frequentist methods that combine p-values. These approaches include Fisher's method and an extension of the Adaptive Rank Truncated Product test (ARTP). The ARTP method adapted to gene- and pathway-environment interaction gives promising results and has been wrapped to R package PIGE [3]. Performances of permutation and parametric bootstrap resampling methods are compared using simulated data with a binary outcome from a case-control study. An illustration on a real dataset on circadian genes and breast cancer risk will be provided.

References

- [1] Buzkova, P., Lumley, T., & Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of human genetics*, 75(1), 36-45.
- [2] Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1), 1-77.
- [3] Liquet, B., & Riou, J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC medical research methodology*, 13(1), 75.

GADAG : un paquet R dédié à l'inférence de Graphes Acycliques Dirigés par maximum de vraisemblance pénalisé

M. Champion^a, V. Picheny^b and M. Vignes^c

^a Laboratoire MAP5, Université Paris Descartes
45 rue des Saints-Pères, 75270 Paris cedex 06
magali.champion@parisdescartes.fr

^b Unité MIAT, Université de Toulouse, INRA
24 chemin de borde-rouge, 31326 Castanet-Tolosan cedex
victor.picheny@toulouse.inra.fr

^c Institute of Fundamental Sciences, Massey University
Palmerston North, New Zealand
m.vignes@massey.ac.nz

Mots-clefs Inférence de réseaux, Statistiques, Optimisation convexe.

Introduction

L'apprentissage automatique de graphes est une problématique de recherche très active ces dernières années, permettant notamment de comprendre les interactions existant entre différentes entités d'un même système complexe. En biologie, la reconstruction de réseaux de régulation de gènes permet par exemple d'identifier les mécanismes de régulation (activation ou inhibition) de l'expression des gènes à partir d'observations de données d'expression de ces mêmes gènes. Elle fait naturellement écho aux enjeux statistiques de grande dimension (grand nombre de variables pour un petit nombre d'échantillons seulement) et de parcimonie, puisque l'on cherche uniquement les interactions principales par souci de qualité et d'interprétabilité de celles-ci.

Méthode d'estimation

Une modélisation classique des réseaux qui est largement utilisée dans la littérature consiste à supposer que les nœuds du graphe sont liés linéairement les uns aux autres [1]. Afin d'introduire de la causalité dans le modèle, seuls les graphes acycliques dirigés (DAGs) sont alors considérés. Dans le cadre de la pré-publication [2], nous avons montré qu'un bon DAG candidat était solution d'un problème de maximum de vraisemblance pénalisé. La difficulté principale de ce problème d'estimation réside dans l'exploration de l'ensemble des solutions (ensemble des DAGs), qui est NP-complexe [3].

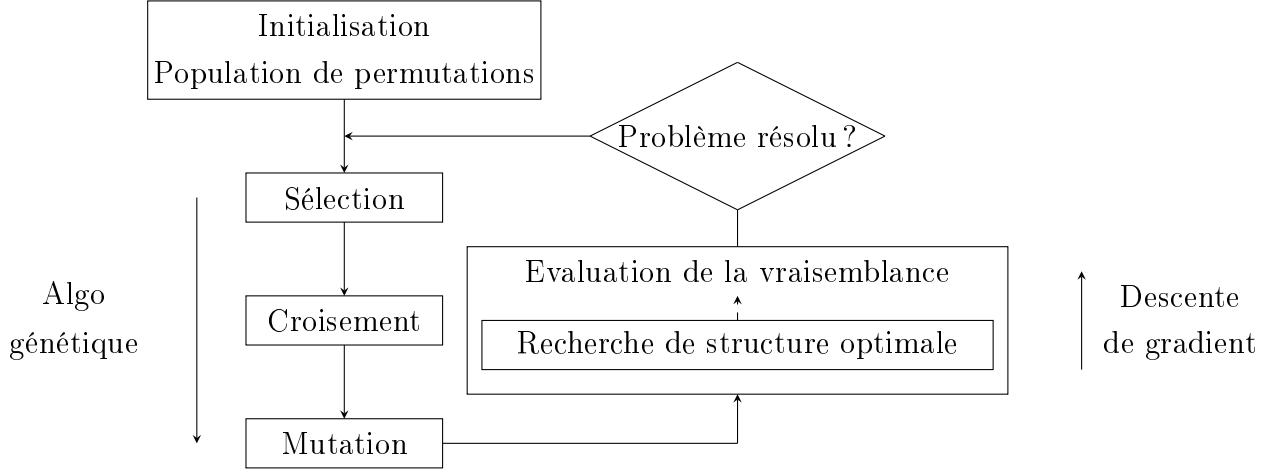
Algorithme GADAG

L'algorithme GADAG, que nous avons développé et publié récemment sur le CRAN, propose une manière originale dédiée à la résolution de ce problème d'inférence. Il est principalement basé sur une décomposition spécifique des DAGs, qui permet de réécrire le problème de maximum de vraisemblance pénalisé sous la forme de deux-sous problèmes de minimisation :

1. la recherche d'un ordre topologique entre les nœuds du graphe suivant le nombre d'arêtes entrantes, modélisé sous la forme d'une permutation,
2. la recherche de la structure du graphe en elle-même (nombre d'arêtes).

Cette reparamétrisation du problème de base en simplifie considérablement la résolution : trouver le nombre d'arêtes du graphe à ordre des nœuds du graphe connu se ramène ainsi à une classe de problèmes classiques de minimisation de fonctions convexes sous contraintes convexes,

que l'on peut résoudre à l'aide d'un algorithme de descente de gradient. Afin d'explorer intelligemment l'espace des permutations, GADAG est basé sur un algorithme génétique [4] : une population de permutations évolue suivant des opérateurs évoquant les processus d'évolution génétique de telle sorte à minimiser la fonction objectif (maximum de vraisemblance pénalisé). Une boucle interne permet d'associer, à chaque permutation explorée, la structure de DAGs optimale, solution du sous-problème de minimisation.



Problèmes techniques

Afin de limiter les temps de calcul, la partie concernant la descente de gradient, qui peut être vue comme une version matricielle du LARS [5], a été codée en Rcpp. De même, l'algorithme génétique est naturellement parallélisé à l'aide des packages `doParallel` et `foreach`. Les opérateurs de l'algorithme génétique ont été définis de manière ad-hoc pour répondre spécifiquement à notre problème.

Résultats

GADAG renvoie à l'utilisateur le DAG qui maximise la vraisemblance du modèle pénalisée. Dans le cas où le graphe qui a servi à générer les données est connu, il calcule la performance de la méthode d'estimation en termes de vrais positifs, faux positifs, vrais négatifs, faux négatifs, précision et rappel. Le paquet propose en outre un nombre conséquent de représentations graphiques : graphe estimé mais également évolution de l'algorithme (fonction objectif et indicateurs de convergence).

Références

- [1] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436-1462.
- [2] Champion, M., Picheny, V. and Vignes, M. (2017). Inferring large graphs using ℓ_1 -penalized likelihood. Preprint (<https://hal.archives-ouvertes.fr/hal-01172745>).
- [3] Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In Learning from data (Fort Lauderdale, FL, 1995), vol. 112 of Lect. Notes Stat. Springer, New York, pp. 121–130.
- [4] Michalewicz, Z. (1994). Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag edition.
- [5] Efron, B., Hastie, T., Johstone, I. and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2) :407-99.

Evaluation du package Heemod et mise en place d'un tableau de synthèse en Shiny pour l'analyse de coût-efficacité.

Yonatan Carranza Alarcon ¹, Benoit Liquet ², Sébastien Marque ¹, Louise Baschet ¹

¹ Capionis, ² Université de Pau et des Pays de L'Adour

{ycarranza.alarcon, sebastien.marque, louise.baschet}@capionis.com

benoit.liquet@univ-pau.fr

Abstract

Dans le champs de l'économie de la santé, l'évaluation coût-efficacité est un élément central. La recherche des outils pour mettre en œuvre ce genre de modèle est devenu un enjeu majeur, car l'utilisation de tableurs (de type Excel) est source d'erreurs et limite la traçabilité et le contrôle qualité. Pour cela, le package *Heemod* est une alternative qui nous a semblé intéressante. Nous nous sommes donc attachés à évaluer et analyser le paquet , ainsi que de développer une plate-forme web pour l'utilisateur final et la communauté. Cette dernière utilise le paquet *Heemod*, et a été implémenté entièrement en R avec *Shiny*.

Mots clefs : Markov multi-état, Coût-efficacité, Médico Economie, Shiny

1 Introduction

A ce jour, les méthodes d'évaluation économique en santé sont devenues, partout dans le monde, un instrument nécessaire pour comparer les stratégies médicales qui seront évaluées par les autorités de santé du pays (HAS en France) afin qu'ils puissent prendre une décision *pertinente* (i.e. prix, remboursement...).

Les modèles d'analyses de décision markovien multi-état sont les plus souvent utilisés pour modéliser ce genre d'études [1], accompagnées des indicateurs médico-économiques (i.e. coût-efficacité, RDCR ¹, ICER²) et d'une population simulée (i.e. la cohorte d'études). Aujourd'hui, ces modèles sont principalement implantés à l'aide d'Excel (i.e. Microsoft Office) afin d'offrir une interactivité et une transparence aux autorités. Néanmoins, comme *C. Williams et al. (2016)* l'explique clairement dans [8], cette approche n'est pas pratique pour de multiples raisons, et la communauté aurait besoin d'un outil plus fiable, plus flexible et plus évolutif.

Le paquet *Heemod*[9] a été développé à l'URC ECO pour résoudre tous ces problèmes. Afin de pouvoir utiliser ce paquet auprès des autorités, notre premier objectif est de pouvoir valider celui-ci avec les calculs et les modélisations faites classiquement (avec un tableur), dans le but d'assurer la transparence de ces résultats et de promouvoir un nouvel outil certifié pour l'utilisation des autorités.

Après avoir validé le paquet *Heemod*, nous mettrons à disposition une application web en *Shiny* [7] afin d'abstraire tout la programmation R à de simples interfaces d'entrées et de sorties (i.e. tableau de bord et de synthèse). La suite du travail est structurée par un rappel de modèles markoviens multi-états, suivi par l'évaluation et validation du paquet *Heemod*, puis les interfaces web implémentées en *Shiny*, et enfin la conclusion.

2 Modèle multi-état

Le parcours du patient au travers de l'évolution de la pathologie est modélisée par différents états de santé exhaustifs et mutuellement exclusifs que le patient peut connaître, tels que "Bonne santé", "Malade" et "Mort". Ces

¹ratios différentiels coût-résultat

²Incremental Cost-Effectiveness Ratio

états de santé peuvent être modélisées par une chaîne de Markov³ $X(t)$ à état discret $S = \{1, 2, \dots, r\}$ et à temps continu $t \in \mathbf{R}^+$ (i.e. Markov Multi-état) et leurs intensités de transition sont définies par:

$$\alpha_{hj}(t; \mathcal{F}_{t-}) = \lim_{\Delta t \rightarrow 0^+} \frac{P_{hj}(t, t + \Delta t | \mathcal{F}_{t-})}{\Delta t}, h, j \in S$$

où $P_{hj}(t, t + \Delta) = P(X(t + \Delta) = j | X(t) = h, \mathcal{F}_{t-})$ représentent les probabilités de transition et \mathcal{F} est la filtration de la chaîne de markov $X(t)$ ou “*histoire du processus*” (i.e. l’historique clinique du patient) représentant l’ensemble des événements observés à l’instant t [3]. La figure 1 représente la simulation d’une cohorte dans deux temps différents, nous pouvons constater que certains patients changent d’état pendant la simulation. Les coûts et ef-

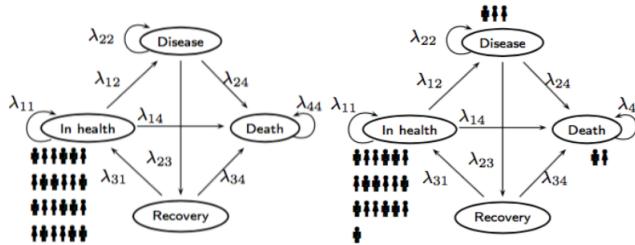


Figure 1: *Simulation Makov Multi-état* [5]. À gauche dans le temps $t = t_0$ (début), et à droite dans un temps t_1 .

ficacités (en termes de QALY principalement, c'est-à-dire années de vie ajustées sur la qualité de vie) sont calculés à chaque itération pendant la simulation et additionnés à la fin, sur l'horizon temporel défini. Enfin, nous utilisons l'indicateur économique *ICER* afin de comparer les stratégies.

3 Évaluation et Validation

L’évaluation du package est réalisée en deux étapes: (1) La première étape a été nommée **Validation interne**, autrement dit, nous avons comparé les résultats de cas réels cliniques (i.e. [6], [4], [2], et autres) implémentés à l'aide de tableur Excel avec les résultats issus des fonctions du paquet *Heemod*. (2) La

³Le caractère de cette chaîne peut être *homogène* ou *non* ainsi que semi-homogène.

deuxième étape a été nommée **Validation externe**, autrement dit, nous avons comparé les résultats obtenus avec le paquet *Heemod* contre notre propre implémentation (i.e. la simulation markovienne multi-état avec les paquets existants tels que *Mstate*, *SemiMarkov*, et autres, et en ajoutant, une couche avec les indicateurs médico-économique). Cette simulation comprend notamment des valeurs extrêmes afin d’évaluer la robustesse du paquet à des situations extrêmes. Nous avons également essayé de modéliser tous les cas possibles et particuliers qui peuvent se présenter lors de cas réels.

4 Tableau de synthèse

L’objectif de créer une application web est de pouvoir fournir une interface web ergonomique pour l’utilisateur final, ainsi qu’une architecture flexible et évolutive (i.e. avec les principes de GRAPS⁴) pour la communauté. Nous avons donc été inspiré par l’architecture reconnue *AngularJS* et *BlurAdmin*⁵ en utilisant *Shiny* comme base, et en l’intégrant avec les librairies *Shiny Dashboard*⁶ et *Shiny Routes*⁷.

Grâce à cette architecture, le développement d’un module est plus simple et le couplage faible à l’application (i.e. le module peut être implanté et testé indépendamment de l’application). Dans la figure 2 (Annexe A.1), nous avons représenté les courbes de survies par stratégie ainsi que l’évolution de la cohorte pendant la simulation.

5 Conclusion

En ignorant toute complexité mathématique et technique, pour la mise en pratique de l’évaluation économique en santé, nous avons pu mettre en oeuvre une version précoce de la plateforme web baptisée *WeMEco*.

Des nouveaux sous-modules de synthèse pourront être ajoutés par des tiers, ainsi que de

⁴General responsibility assignment software patterns (GRAPS)

⁵<http://akveo.github.io/blur-admin/>

⁶<https://rstudio.github.io/shinydashboard/>

⁷<https://apppsilon.github.io/shiny.router/>

nouveaux types de modélisation mathématique (e.g. modèle de markov caché) en mettant à jour le package *Heemod*, et en l'intégrant à travers d'une interface sur *WeMEco*.

References

- [1] Briggs et al. “An Introduction to Markov Modelling for Economic Evaluation”. In: *PharmacoEconomics* 13.4 (1998), pp. 397–409. ISSN: 1179-2027. DOI: 10.2165/00019053-199813040-00003. URL: <http://dx.doi.org/10.2165/00019053-199813040-00003>.
- [2] Matthias Bischof et al. “Cost-Effectiveness of Drug-Eluting Stents in a US Medicare Setting: A Cost-Utility Analysis with 3-Year Clinical Follow-Up Data from the 649”. In: *PharmacoEconomics* 12.45 (2009). DOI: 10.1111/j.1524-4733.2009.00676.x.
- [3] Benoit Liquet. “HDR - Modélisation Statistique et Applications Biomédicales”. In: (2009).
- [4] Joshua A. Ray et al. “An Evaluation of the Cost-Effectiveness of Rituximab in Combination with Chemotherapy for the First-Line Treatment of Follicular Non-Hodgkin’s Lymphoma in the UK”. In: *PharmacoEconomics* 13.4 (2010), pp. 346–357. DOI: 10.1111/j.1524-4733.2009.00676.x.
- [5] G. Baio. *Bayesian Methods in Health Economics*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, 2012. ISBN: 9781439895559. URL: <https://books.google.fr/books?id=m731%5CjXLnCsC>.
- [6] L. Baschet et al. “Cost-effectiveness of drug-eluting stents versus bare-metal stents in patients undergoing percutaneous coronary intervention”. In: *Open Heart* (2016). DOI: 10.1136/openhrt-2016-000445. URL: <http://dx.doi.org/10.2165/00019053-199813040-00003>.
- [7] Chang W et al. “Shiny: Web Application Framework for R.” In: (2016). R package version 1.0.2. URL: <https://CRAN.R-project.org/package=shiny>.
- [8] Claire Williams et al. “Cost-effectiveness analysis in R using a multi-state modelling survival analysis framework: a tutorial”. In: *Medical Decision Making* (2016). DOI: 10.1177/0272989X16651869. URL: <http://eprints.gla.ac.uk/116523/>.
- [9] Antoine Filipovié-Pierucci, K. Zarca, and I. Durand-Zaleski. “Markov Models for Health Economic Evaluation: The R Package heemod”. In: *ArXiv e-prints* (Feb. 2017). R package version 0.9.0. eprint: 1702.03252 (stat.AP). URL: <https://pierucci.org/heemod>.

Annexe

A Interfaces du tableau de bord

A.1 Courbes de survie

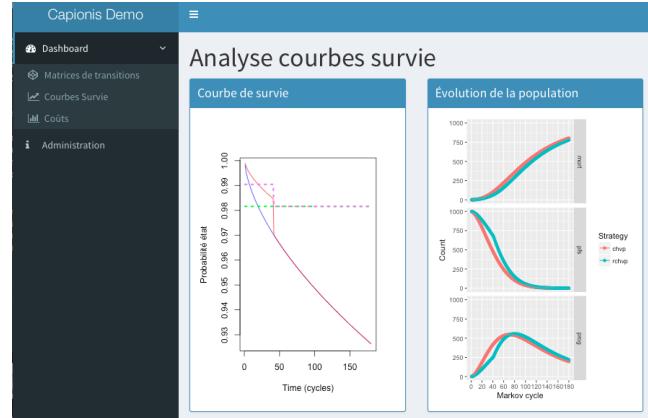


Figure 2: Courbe de survie et Courbe d'évolutions de la cohorte par état et pendant la simulation.

A.2 Matrices de transitions

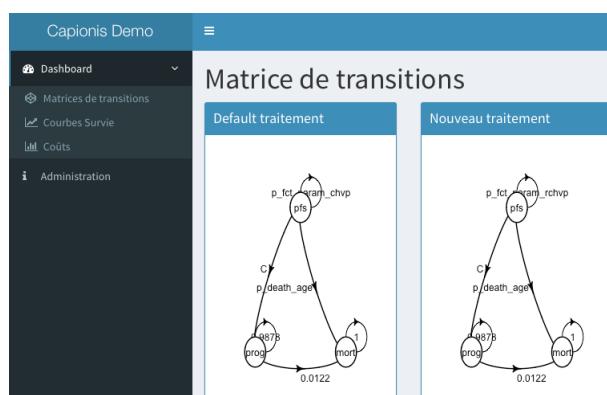


Figure 3: *Tableau de bord - Matrices de transitions des 2 stratégies.*

Approche bloc en ACP group-sparse: le package sparsePCA

Marie Chavent^{a,b} et Guy Chavent^c

^aIMB, UMR 5251, Université de Bordeaux, 33400 Talence

^bInria Bordeaux Sud-Ouest, équipe CQFD, 33405 Talence

marie.chavent@u-bordeaux.fr

^cInria-Paris, 75589 Paris

guy.chavent@inria.fr

Mots clefs : ACP sparse, sélection de variables groupe, optimisation en bloc.

La plupart des algorithmes développés ces dernières années en ACP sparse visent à déterminer une seule composante principale et utilisent le processus de déflation hérité de l'ACP non sparse lorsqu'il s'agit de calculer les composantes suivantes (voir par exemple [2], [5]). Cependant, l'utilisation de la déflation en ACP sparse où les composantes et les loadings ne sont plus nécessairement orthogonaux peut mener à des difficultés [4] et on peut s'attendre à ce que l'optimisation directe de la variance selon tous les loadings soit plus efficace qu'une approche séquentielle. Certains auteurs ont ainsi proposés des algorithmes déterminant tous les loadings simultanément : Zou et al. [6] résolvent l'ACP sparse comme un problème de type régression alternée et Journée et al. [3] utilisent une approche duale de type bloc.

Nous allons présenter dans cette communication une approche de type bloc en ACP group-sparse où les variables sont organisées en groupes et où la sparsité s'applique aux groupes de variables et non plus aux variables individuelles. Cette approche généralise au cas de variables groupe la formulation ℓ_1 -sparse de Journée et al. [3] et l'algorithme associé.

Ce nouvel algorithme d'ACP group-sparse a été implémenté dans le package R **sparsePCA** qui est disponible à l'adresse suivante : <https://github.com/chavent/sparsePCA>. Nous présenterons les principales fonctionnalités de ce package et nous comparerons sur une étude de simulation les approches bloc et déflation en ACP group-sparse [1]. Cette étude de simulation montrera également comment la prise en compte de l'information sur les groupes en ACP group-sparse peut permettre de mieux retrouver les patterns de sparsité qu'avec l'ACP sparse.

Références

- [1] Chavent, M., Chavent, G. (2017). Group-sparse bloc PCA and explained variance, *arXiv preprint arXiv:1705.00461 [stat.ML]*
- [2] D'Aspremont, A., Bach, F., El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(Jul):1269-1294.
- [3] Journée, M., Nesterov, Y., Richtarik, P., Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517-553.
- [4] Mackey , L.W. (2009). Deflation methods for sparse pca. In *Advances in neural information processing systems*, 1017-1024.
- [5] Shen, H., Huang, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015-1034, 2008.
- [6] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265-286.

productivity : un package dédié aux calculs d'indices de productivité par la méthode DEA

K H. Dakpo^a, Y. Desjeux^b and L. Latruffe^c

^a UMR Economie Publique
INRA
Avenue Lucien Brétignières
78850 Thiverval-Grignon
k-herve.dakpo@inra.fr

^b UMR SMART-LEREKO
INRA
4, allée A.Bobierre
35011 Rennes CEDEX
yann.desjeux@inra.fr

^c UMR SMART-LEREKO
INRA
4, allée A.Bobierre
35011 Rennes CEDEX
laure.latruffe@inra.fr

Mots clefs : productivité ; indice de Färe-Primont ; indice de Lowe ; indice de Malmquist.

La productivité est un facteur primordial de l'amélioration de la compétitivité et des revenus, et plus généralement de l'analyse et l'évaluation des politiques, qu'elles soient publiques ou privées. Un grand pan de la littérature en sciences économiques s'est intéressé et s'intéresse toujours aux mesures de la productivité d'entités, qu'il s'agisse d'entreprises ou d'organisations évoluant dans divers domaines d'activité, des secteurs d'activité ou des pays. De façon générale dans le domaine l'économie de la production, les processus de production impliquent l'utilisation de plusieurs facteurs de production ou « inputs » (par exemple, la main d'œuvre, le capital) pour créer plusieurs produits ou « outputs ». La productivité totale des facteurs de production est alors définie comme le ratio d'un output unique « global » (c'est-à-dire agrégeant tous les outputs produits) sur un input unique « global » (agrégeant tous les inputs utilisés). Cette mesure fait consensus au sein de la communauté scientifique. Son avantage est qu'elle considère les systèmes de production dans leur ensemble, contrairement aux mesures partielles de la productivité (ratio d'un output spécifique sur un input spécifique) qui ne traitent que d'un input et d'un output. Néanmoins, le calcul de la productivité totale des facteurs est complexe. En effet, l'agrégation des outputs et des inputs requiert l'utilisation de pondérations ou de poids spécifiques, afin de pouvoir agréger des inputs (ou outputs) ayant des unités de mesure différentes.

Dans la littérature plusieurs méthodes, ou indices, de productivité totale existent. Par exemple, les indices de Laspeyres, de Paasche, de Fisher ou encore de Törnqvist utilisent les prix comme facteurs d'agrégation et restent très largement utilisés par les instituts de statistiques. D'autres indices moins utilisés, comme l'indice de Malmquist ou l'indice de Hicks-Moorsteen, présentent l'avantage de ne pas nécessiter d'information sur les prix observés. Ces deux indices utilisent en effet des prix implicites, c'est-à-dire estimés, pour effectuer l'agrégation.

Un indicateur (ou indice) de productivité fiable doit vérifier un certain nombre d'axiomes et de tests (voir Diewert [1] pour une liste plus ou moins exhaustive). Toutefois, malgré leur popularité, aucun

des indices ci-dessus mentionnés ne vérifie la propriété de transitivité, qui permet de réaliser de façon robuste des comparaisons de productivité multilatérales (c'est-à-dire entre entreprises ou pays) et multi-temporelles (c'est-à-dire entre années). En d'autres termes les indices cités ci-dessus ne sont utiles que pour des comparaisons deux par deux. Récemment, les indices de Färe-Primont et de Lowe ont été proposés par O'Donnell [2], O'Donnell [3] afin de palier au problème de la transitivité sus-évoqué. Alors que le premier indice ne nécessite pas les prix, le deuxième utilise quant à lui les prix comme facteurs d'agrégation.

De façon pratique, l'indice de Malmquist peut être calculé sous le logiciel R à l'aide du package **FEAR** (Wilson [4]). Ce package n'est cependant pas publié dans le CRAN, et les dernières versions requièrent une licence. Le package **nonparaeff** [5] permet également le calcul de l'indice de Malmquist ainsi que sa décomposition. En effet, les indices de productivité mesurent les changements dans la productivité d'une entité de production au cours du temps, et ces changements peuvent être dus au progrès technologique ou à des changements dans l'efficacité de l'entité. La décomposition d'un indice permet alors de mesurer ces deux sources de changements. Enfin, en ce qui concerne les indices nécessitant des prix tels que les indices de Laspeyres ou de Fisher, nous n'avons trouvé aucun package sous R permettant leur calcul et décomposition.

C'est dans ce contexte que nous proposons le package **productivity** qui permet la mesure et la décomposition de l'indice de Malmquist, mais aussi des indices transitifs robustes de Färe-Primont et de Lowe afin notamment de permettre une comparaison avec des indices classiques. En plus de la décomposition en progrès technique et changement d'efficacité, ce package permet de décomposer le changement d'efficacité en changement de l'efficacité technique, changement de l'efficacité d'échelle et changement de l'efficacité mixte. De plus, ce package va plus loin que les packages existants **FEAR** et **nonparaeff** en terme de décomposition, puisque dans **productivity** la composante progrès technologique de l'indice de Malmquist est décomposée en une partie neutre (c'est-à-dire que le progrès est intervenu de manière équivalente pour tous les inputs et les outputs), et une partie non-neutre (c'est-à-dire que le progrès est biaisé vers la réduction d'inputs spécifiques ou l'augmentation d'outputs spécifiques). Sur la base de fonctionnalités proposées par les packages **doParallel** et **foreach** [6], le package **productivity** présente également l'avantage de proposer des possibilités de calcul parallèle paramétrables afin d'optimiser les temps de calcul des différents indices de productivité proposés.

Références

- [1] Diewert, W.E. (1992). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3:211-48.
- [2] O'Donnell, C.J. (2011). The sources of productivity change in the manufacturing sectors of the US economy. *Working Papers WP07/2011: School of Economics, University of Queensland, Australia*.
- [3] O'Donnell, C.J. (2012). Nonparametric Estimates of the Components of Productivity and Profitability Change in U.S. Agriculture. *American Journal of Agricultural Economics*, 94:873-90.
- [4] Wilson, P.W. (2008) FEAR: A software package for frontier efficiency analysis with R. *Socio-Economic Planning Sciences*, 42:247-54.
- [5] Oh, D.-H., Suh, D. (2013). Package ‘nonparaeff’: Nonparametric Methods for Measuring Efficiency and Productivity. <https://CRAN.R-project.org/package=nonparaeff>.
- [6] Weston, S., Calaway, R. (2015). Getting Started with doParallel and foreach. Available on <https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstartedParallel.pdf>.

Bassins d'habitat pour les Pyrénées-Atlantiques

E. Damitio^a and D. Sauvignet^b

^a Observation / Statisticienne,
AUDAP
Petite Caserne
2 allée des Platanes - BP628
64106 BAYONNE
e.damitio@audap.org

^b Observation / Habitat, Sociologue
AUDAP
Petite Caserne
2 allée des Platanes - BP628
64106 BAYONNE
deborah.sauvignet@audap.org

Mots clefs : Urbanisme, Socio-démographie, Statistique, Flux.

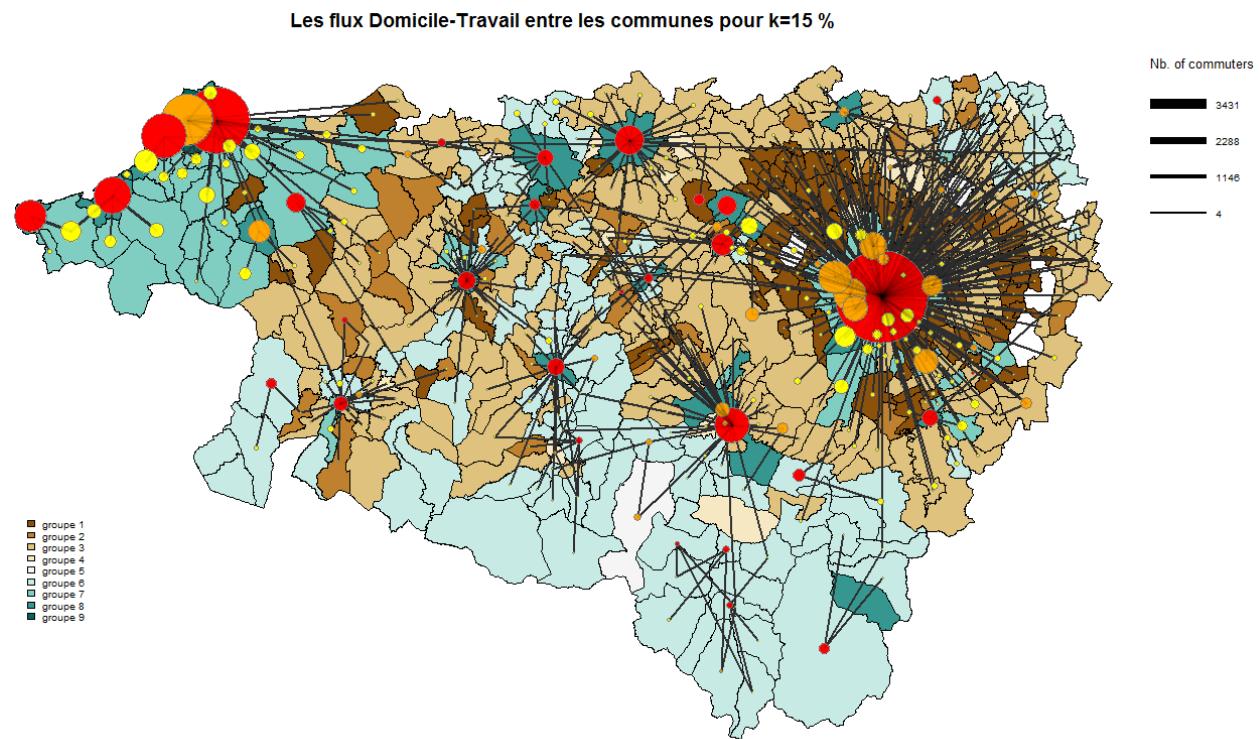
L'utilisation de R dans le champ de l'urbanisme n'est pas usuelle. Dans cet exercice nous n'avons pas effectué de programmation mais nous avons combiné différentes méthodes statistiques et travaillé sur la représentation de l'information sous forme cartographique.

L'objectif est d'identifier les différents bassins d'habitat du département des Pyrénées-Atlantiques. Un bassin d'habitat peut se définir comme une « aire de chalandise » résidentielles autour de pôles d'emploi généralement urbains [1].

La méthodologie se décline en trois étapes :

1. Création d'une base de données de 26 variables quantitatives issues de différentes sources (INSEE, Etat Civil, DGFIP, SIRS, SOES et SAFER) pour caractériser les 547 communes des Pyrénées-Atlantiques [2]. Elle regroupe des informations démographiques, économiques, fiscales et foncières.
2. Travail sous R-Studio : Ces 26 variables nous ont permis de réaliser une Analyse en Composantes Principales pour travailler sur des variables dé-correlées et ainsi faire une classification de 9 classes. La projection des 9 classes sur la carte à l'échelle communale des Pyrénées-Atlantiques avec le package cartography [3] nous permet de spatialiser les communes selon leurs classes.
3. Pour compléter la classification, nous avons utilisé la méthode des flux dominants/dominées [4] avec les données INSEE « Domicile / Travail » grâce au package flows [5]. Cette méthode permet de voir les dynamiques entre les différentes communes du département et d'observer des relations qui existent selon les types de classifications et d'en dessiner des bassins d'habitat.

Cette analyse nous permet de décliner, dans ce département, 14 bassins d'habitat.



Références

- [1] Jean-Claude.Driant (2016). Bassin d'habitat, Politique du logement, analyses et débats, 1 p.
- [2] Direction régionale de l'Environnement, de l'Aménagement et du Logement de Basse-Normandie (2013). Typologie des territoires bas-normands sous l'angle de l'habitat, 14 p.
- [3] Timothée Giraud [cre, aut], Nicolas Lambert [aut] (2016). Create and integrate maps in your R workflow, Package R.
- [4] John D.Nystuen, Michael F.Dacey (1961). A graph theory interpretation of nodal regions, p 29–42.
- [5] Timothée Giraud [cre, aut], Laurent Beauguitte [aut], Marianne Guérois [ctb] (2016). Selections on flow matrices, statistics on selected flows, map and graph visualisations, Package R.

Analyse des débats sur le secret bancaire au Parlement suisse avec R

Marion Deville *a*, Emmanuel Rousseaux *b*, Pierre-Alexandre Fonta *c*, Roy Gava *d* et Yannick Rochat *e*

a Institut de Recherches Sociologiques, Université de Genève, 40 bd du Pont d'Arve, 1211 Genève

Marion.Deville@unige.ch

b Institut de Systèmes d'Informations, Université de Genève, 40 bd du Pont d'Arve, 1211 Genève

Emmanuel.Rousseaux@unige.ch

c Pôle national LIVES, Université de Genève, 40 bd du Pont d'Arve, 1211 Genève

Pierre-Alexandre.Fonta@unige.ch

d Département de science politique et relations internationales, Université de Genève, 40 bd du Pont d'Arve

Roy.Gava@unige.ch

e Section des sciences du Langage et de l'Information, Université de Lausanne, UNIL-Dorigny, 1015 Lausanne

yannick.rochat@unil.ch

Mots clefs : parlement, débat, synthèse, réseaux, visualisation, secret bancaire

Ce *Lightning Talk* présente une analyse réalisée avec le logiciel R de l'évolution du débat sur le secret bancaire en Suisse. Sur la base de 20 ans de débats parlementaires aux Chambres fédérales suisses, nous rendons compte de l'évolution de la dynamique entre les groupes parlementaires et acteurs mentionnés dans les débats (OCDE, US, France), révélant ainsi les évolutions dans les coalitions et les aspects sémantiques mobilisés. Les parlements sont des lieux où la place du discours est centrale. Les débats sont retranscrits par les administrations et disponibles publiquement. Ces délibérations en plénière constituent une quantité de matériel d'étude conséquente, difficile à traiter de manière transversale automatiquement. Ce traitement automatique a connu d'importantes avancées avec la Fabrique de la loi, qui traite le texte des interventions dans sa procédure de synthèse des étapes parlementaires. Si ces outils d'organisation des données existent et sont performants, la question se pose aujourd'hui de donner du sens à ces données avec des analyses permettant de retracer les dynamiques de débat.

Les données d'étude ont été extraites et nettoyées avec le logiciel R à partir des textes bruts des débats au format PDF. Le texte extrait a été analysé à l'aide du package **spnet** [1] en suivant la méthode Deville et Rousseaux [2] [3]. Cette méthode synthétise l'information issue des textes des mémoires de manière à faire ressortir des contrastes dans les temps du débat sur les aspects relationnels et sémantiques. La restitution visuelle a pour objectif de permettre une analyse fine et globale du débat en plénière en tenant compte des interactions entre les parlementaires et de la temporalité. Analyser le texte des interventions parlementaires à ce niveau permet de se diriger vers des descriptions nuancées et de tester des hypothèses pour mettre en lumière des mécanismes difficiles à observer depuis les données sous forme textuelle.

Cette présentation est organisée de la manière suivante : (1) Objectifs de l'étude dans le domaine des sciences sociales, (2) présentation du corpus, (3) présentation des scripts d'extraction et de nettoyage du texte, (4) présentation du package **spnet** (5) résultats de l'étude (6) intérêts de l'utilisation de R pour cette analyse, enjeux de réplicabilité, ouvertures sur la recherche appliquée et concernant le regard citoyen sur les institutions.

Références

- [1] Rousseaux, E., Deville, M., Ritschard, G., Package 'spnet', 2014.
- [2] Deville, M., Rousseaux, E., A spatial network approach for measuring the differentiation between content and relational dynamics in the political debate. 2014.
- [3] Deville, M., Rationalités en débat, ISTE, à paraître.

A. Antoni et T. Dhorne

Université de Bretagne Sud
Lab-STICC UMR 6285
8 rue Montaigne, BP 561, 56017 VANNES Cedex
arlette.antoni@univ-ubs.fr
thierry.dhorne@univ-ubs.fr

Mots clefs : Scénarios, Enseignement, Environnement de travail

On présente un outil de gestion scénarios d'étude statistique en R, comme en présente Chatfield [1].

De par ses fonctionnalités, R est susceptible de répondre à des objectifs d'«encadrement» des utilisateurs, étudiants ou professionnels peu compétents.

Nous avons développé des outils dédiés :

1. au développement
2. à la visualisation (et donc à l'apprentissage)
3. à la mise en œuvre
4. au reporting

de scénarios d'analyses R.

1 Scénario d'analyse

Un scénario d'analyse R est une suite de tâches (instructions ou groupes d'instructions) devant être menées dans un ordre donné, avec des contraintes données et validées par un retour ou par une interprétation.

Cette vision est une vision normative qui émane d'une expertise. Le niveau des contraintes et des retours permettent cependant de rendre « flexible » la norme.

Un exemple pour éclairer le propos peut être trouvé dans l' *étude graphique d'une variable statistique continue pour la modélisation*

On peut comme exemple proposer un scénario avec ses différentes tâches :

1. réaliser plusieurs histogrammes (quelles contraintes ?) et faire un diagnostic sur la variable,
2. proposer (éventuellement) une transformation de la variable (pour la « rapprocher » d'un modèle connu) et la réaliser
3. réaliser un graphe qt - qe adapté au modèle et valider ou invalider le modèle,
4. réaliser un test d'adéquation (qui peut être spécifié) et conclure quant au modèle proposé.

Comme on le voit sur cet exemple, une tâche est constituée d'éléments clairement spécifiés (calculs, traitements, graphiques,...) et d'éléments plus ouverts (validation, diagnostic, commentaire,...)

2 Tâches

Une tâche est identifiée par

-
- son libellé,
 - les opérandes, (sur quoi s'applique(ra) la tâche ?)
 - l'opération, (en quoi consiste(ra) la tâche)
 - ses prérequis (tâches antérieures qui doivent avoir été réalisées),
 - son état (réalisé, non réalisé ou inconnu),
 - une restitution associée (validation, commentaire, réponse à des questions).

On présente en détail la tâche « réaliser plusieurs histogrammes et faire un diagnostic sur la variable » :

3 Développement de scénarios

Le développement de scénarios peut être réalisé par un expert (enseignant, spécialiste,...), il a donc alors un aspect normatif ou par un non expert (étudiant, junior,...), il a donc alors un aspect exploratoire.

Les scénarios sont définis formellement comme un graphe ou éventuellement un arbre de tâches. On fournit des outils empruntés ou non à des paquets existants pour spécifier et valider les scénarios de manière interactive ou automatique, ainsi que pour les mettre à disposition.

4 Visualisation de scénarios pour la mise en œuvre et l'apprentissage

La visualisation des scénarios s'appuie sur des outils graphiques préexistants. Ils sont modifiables facilement par l'utilisateur.

5 Mise en œuvre

La mise en œuvre de scénarios dépend fortement du niveau de spécification de ceux-ci :

- pour des scénarios normatifs préétablis, il s'agit essentiellement d'appliquer les outils et méthodes préconisés aux objets d'études,
- pour des scénarios créatifs, une grande liberté est potentiellement laissée à l'utilisateur.

Le retour attendu est constitué (pour le moment ?) :

- de code R éventuellement commenté
- de retours sous forme de questions ouvertes ou fermées

6 Reporting

Le reporting peut être orienté vers :

- une diffusion classique sous forme de notes ou de rapports lorsque l'étude réalisée est balisée
- une évaluation (par le pédagogue ou le responsable) lorsque l'étude est un exercice ou revêt un caractère expérimental.

La diffusion classique est gérée par l'intermédiaire de paquets dédiés (knitr, Rmarkdown,...). L'évaluation est réalisée sur le code par divers tests et sur les retours numériquement ou par quelques traitements textuels.

Références

Chatfield, C. (1995). Problem Solving. A statistician's guide. Second Edition. *Chapman and Hall/CRC - London*

Analyse spatiale multi-échelles de données écologiques avec adespatial

S. Dray^a

^a Laboratoire de Biométrie et Biologie Evolutive
CNRS
Université Lyon 1

Mots clefs : Statistique Spatiale, Analyses Multivariées, Ecologie.

Dans de nombreux domaines, les questions thématiques conduisent à analyser des données multivariées et géoréférencées. En écologie, par exemple, l'étude des patrons de variation de la biodiversité conduit à analyser simultanément les distributions d'espèces dans le paysage afin d'identifier les principales structures spatiales en abordant notamment les notions d'autocorrélation et d'échelle.

Dans ce cadre, le package ‘adespatial’ propose de nombreux outils pour l’analyse spatiale multi-échelles de données multivariées. La package s’appuie sur un cadre théorique basé sur l’utilisation de matrices de voisinage afin d’intégrer l’espace en analyse de données multivariées. La diagonalisation de ces matrices permet d’aborder la notion d’échelle spatiale en fournissant des bases orthogonales de prédicteurs spatiaux. Du point de vue de l’implémentation, le package ‘adespatial’ se place à l’interface de plusieurs packages. L’information spatiale est codée sous la forme de graphe de voisinage grâce aux fonctions fournies par le package spdep. Les fonctionnalités relatives à l’aspect multivarié s’appuie sur le package ade4.

O. Eterradossi^a

^aPôle Recherche sur les Interactions des Matériaux avec leur Environnement
Centre des Matériaux des mines d'Alès (Institut Mines -Telecom)
EMA, Hélioparc, 2 av. P. Angot, F-64000 PAU (France)
olivier.etterradossi@mines-ales.fr

Mots clefs : Visualisation, Psychophysique, Physique, Couleur, Colorisation

A la fois propriété du monde qui nous entoure, perception, concept cognitivo-culturel et moyen de communication, “la couleur” nous est si familière qu’on l’utilise de manière presque inconsciente. Son utilisation dans **R** est cependant moins triviale qu’il n’y paraît car elle peut y être soit une donnée à traiter soit un moyen d’ajouter une dimension supplémentaire à un graphique.

L’objectif de ce survol est de faire un point sur la manière dont **R** traite ces deux aspects et d’examiner quelques problèmes pouvant survenir à leurs points de tangence. Pour ce faire, la présentation explorera trois sujets (les deux premiers rapidement et le dernier plus en détail) :

Les **packages dédiés aux approches physiques**, qui manipulent des données vectorielles ou matricielles qui peuvent être importés d’instruments de mesure ou simulées. Si ce type de données est principalement exploité pour de l’identification, de la quantification ou de la classification, il nous intéressera ici en tant que base de l’expression de la couleur dans des espaces colorimétriques physico-réalistes.

Les **packages psychophysiques**, qui sont principalement dédiés à l’étude des relations entre stimuli et perceptions mais avec des conséquences en matière d’utilisation de la couleur (comme le développement de palettes à propriétés colorimétriques controlées ou compatibles avec les déficiences visuelles, ces dernières intéressantes en matière de rédaction des publications par exemple).

Les **fonctions dédiées à la colorisation des graphiques**, qui sont dispersées dans une multitude de packages complémentaires, concurrents ou redondants. On adoptera pour les survoler une approche pragmatique basée à la fois sur la typologie des outils (la grammaire des palettes) et sur les buts poursuivis. Pour ce faire, on mettra évidemment d’abord en avant les packages incontournables et les fonctions de base, mais on montrera aussi qu’il existe des fonctions astucieuses enfouies dans des packages parfois inattendus. On montrera également quelques développements spécifiques.

Références

- [1] Zeileis, A., Hornik, K., Murrell, P. (2009). Escaping RGBland: selecting colors for statistical graphics. *Computational Statistics and Data Analysis*, **53**, 3259-3270.
- [2] Brewer, C.A. (1994). Color Use Guidelines for Mapping and Visualization. in *Visualization in Modern Cartography*, MacEachren A.M. et Taylor D.R.F. eds, Elsevier Science, 123-147
- [3] Maloney, L.T., Yang, J.N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, **3**(8), 573-584

Efficient simulation of complex queueing systems with the R package queuecomputer

Anthony Ebert a, Paul Wu b, Kerrie Mengersen c and Fabrizio Ruggeri d

a School of Mathematical Sciences

A-ARC Centre of Excellence in Mathematical and Statistical frontiers (ACEMS), Queensland University of Technology

A-School of Mathematical Sciences, QUT, GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA

ac.ebert@qut.edu.au

b School of Mathematical Sciences

B-ARC Centre of Excellence in Mathematical and Statistical frontiers (ACEMS), Queensland University of Technology

B-School of Mathematical Sciences, QUT, GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA

p.wu@qut.edu.au

c School of Mathematical Sciences

C-ARC Centre of Excellence in Mathematical and Statistical frontiers (ACEMS), Queensland University of Technology

C-School of Mathematical Sciences, QUT, GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA

k.mengersen@qut.edu.au

d School of Mathematical Sciences

D-ARC Centre of Excellence in Mathematical and Statistical frontiers (ACEMS), Queensland University of Technology

D-School of Mathematical Sciences, QUT, GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA

D-Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI-CNR)

D-Via Alfonso Corti, 12 20133 Milano - Italy

fabrizio@mi.imati.cnr.it

Mots clefs : queues, queueing theory, discrete event simulation, operations research

Résumé (Anglais)

The R package `queuecomputer` [1] implements a new algorithm for queueing system simulation. Speedups of more than 2 orders of magnitude are observed compared to existing discrete event simulation (DES) packages `simmer` [2] and `simpy` [3]. With the example of an international airport terminal, we show how to use this package to model complex queueing systems within R. We also demonstrate how to embed a queueing simulation within an approximate Bayesian computation (ABC) algorithm.

Références

- [1] Ebert, A., Wu, P., Mengersen, K., Ruggeri, F. (2017). Computationally Efficient Simulation of Queues: The R Package `queuecomputer`. arXiv preprint arXiv:1703.02151
- [2] Ucar, I., Smeets, B. (2016). `simmer`: Discrete-Event Simulation for R. R package version 3.5.1, URL <https://CRAN.R-project.org/package=simmer>.
- [3] Lünsdorf, O., Scherfke, S. (2013). `simpy`. python package version 3.0.10, URL <https://simpy.readthedocs.io/en/latest/index.html>.

Colin FAY

ThinkR
63 rue de la Pigaci  re
14000 Caen

colin@thinkr.fr

Mots clefs : API, web, base de données.

À l'heure actuelle, les API sont partout. Réseaux sociaux, Open Data, applications, services, base de données... difficile d'avoir une vue exhaustive sur le nombre exact de ces interfaces. À titre d'exemple, en ce début d'année 2017, ce ne sont pas moins de 17000+ API qui sont recensées sur Programmable Web, le site de référence sur les API.

Pourtant, et pour paraphraser H. Wickhama qui lui même paraphrase Tolstoï : « *Good API are all alike, but every messy API is messy in its own way* ». Car oui : là où les APIs populaires peuvent tendre vers une structure simple, bien documentée et presque universelle, d'autres API peuvent présenter des spécificités parfois difficiles à saisir. Des particularités qui peuvent se révéler un variable casse-tête pour les utilisateurs de R qui souhaitent interroger ces plateformes. Comment accéder à ces données simplement, sans perdre des heures de recherche et d'erreur ?

À travers une série d'exemples, Colin parcourra une partie de l'environnement des API web, partant d'interfaces documentées et accessibles, pour se diriger vers les coins les plus obscurs du web programmable. Au menu: un workflow basique, quelques bonnes pratiques, et un retour d'expérience sur des heures d'essais-erreurs à interroger ces interfaces.

Colin FAY

ThinkR
63 rue de la Pigaci  re
14000 Caen

colin@thinkr.fr

Mots clefs : Visualisation de donn  es, datajournalisme.

L'omnipr  ence moderne des donn  es a rendu indispensable la croissance d'une discipline bien sp  cifique : celle de la datavisualisation. Un essor qui a d'ailleurs permis l'  mergence de nouveaux m  tiers, comme celui de datajournalisme ou de data designer.

Aujourd'hui, rapports, articles scientifiques ou de presse ´ecrite, web ou t  l  ... les visualisations de donn  es sont partout, apportant des informations qui peuvent  tre strat  giques dans les prises de d  cisions citoyenne, politique, ou encore ´conomique. Pourtant, et l   o   on pourrait croire que "les donn  es sont objectives", une repr  sentation graphique est loin d' tre neutre : bien au contraire, elle rel  ve (presque) toujours d'un choix ´ditorial, d'une volont  de passer un message sp  cifique, de faire parler les donn  es d'une certaine voix.

Le message de ce lightning talk ? Il est possible de faire dire beaucoup de choses ´ un m  me jeu de donn  es — c'est ce que Colin viendra d  montrer, en prenant comme source un dataset en Open Data, pour le d  cliner en dix visualisations diff  rentes avec R.

C Genolini^{a,b}

^aUniversité Paris-Nanterre
^bZébrys
cgenolin@u-paris10.fr

Mots clefs : Statistique, Interface Homme Machine, GUI, data management

1 Introduction

1.1 *R++, the Next Step*

R++, the Next Step est un projet de développement d'une nouvelle implémentation de R. Il a pour vocation d'être compilable, d'intégrer en natif la gestion du parallélisme et de permettre l'exploitation des bases de données de grande dimension. Mais surtout, *R++, the Next Step* est intégré dans une interface homme machine moderne et conviviale, spécifiquement conçue pour les analyses statistiques.

1.2 Interaction Homme Machine

L'Interaction Homme Machine est la science ayant pour objectif d'étudier la manière dont les humains interagissent avec les ordinateurs afin d'ensuite concevoir des outils plus ergonomiques. Pour cela, des séances de brainstorming réunissant utilisateurs (dans notre cas les statisticiens) et informaticiens sont organisées. Dans un premier temps, l'objectif est de définir les tâches qui sont particulièrement ardues, pénibles à réaliser ou a fort risque d'erreur : tout ce qui fait cauchemarder les statisticiens. Ensuite, des solutions sont collectivement imaginées. Enfin, un prototype vidéo, illustration par l'exemple du problème et de sa solution, est élaboré.

Dans le cas présent, la "tâche ardue" identifiée a été le data management (ou nettoyage des données).

2 Le data management

La collecte des données est une opération longue et délicate, parsemée d'embûches... qui au final conduit souvent à la constitution d'un fichier de données comportant des erreurs ou des éléments non exploitables. Le data management consiste à corriger les erreurs que l'on trouve dans les données brutes afin de les préparer à l'analyse. Les méthodes de nettoyage sont assez variées : détection et correction des valeurs aberrantes, vérification que le logiciel a correctement typé les colonnes, élimination des doublons... C'est une tâche généralement fastidieuse. Elle prend, selon les sources, entre 30% et 50% du temps global consacré à une analyse statistique.

2.1 Les cauchemards du data management

Lors de séances de prototypage vidéo, nous avons identifiés différents aspects du data management particulièrement problématiques :

- Détection des valeurs aberrantes : dans toutes les études, on trouve des étudiants qui ont 180 ans, ou qui mesurent 180m... La vérification doit se faire colonne par colonne. En outre, la position exacte d'une valeur aberrante peut être compliquée à déterminer, il est donc difficile de la corriger.

- Le typage erroné : une colonne d'entier qui contient la valeur 2O (le chiffre 2 puis la lettre O) sera identifié comme une colonne factor. Elle sera donc considérée comme un factor dans un summary ou dans une régression linéaire.
- Homme, HOMME, homme et H seront considérées comme des modalités différentes. Identifier toutes les différentes versions d'une modalité puis les fusionner peut prendre du temps.
- Les variables ordonnées sont systématiquement considérées comme des factors. Il faut corriger à la main.

2.2 Une IHM dédiée

Dans cette présentation, nous vous proposons de découvrir un nouvel axe de l'Interaction Homme Machine dédié spécifiquement au data management. Pour chacun des points listés ci-dessus, nous avons imaginé différentes solutions. Elles ont été présentées aux utilisateurs. Ils ont choisi ce qui paraissait le plus adapté à leurs besoins. Nous avons ensuite compilé l'ensemble des solutions dans une Interface Homme Machine dédié au data management. C'est cette interface que nous vous présentons aujourd'hui.

Pour que le data management devienne un plaisir...

M. Gousseff^a

^aChaire Décisionnel Connaissance Client
Fondation Université Bretagne Sud
8 rue Montaigne, 56017 Vannes
matthieu.gousseff@univ-ubs.fr

Mots clefs : Clustering, ACM, Indice de Rand.

Introduction

La classification non-supervisée, ou partitionnement, est fréquemment utilisée à des fins de segmentation ou de typologies de clientèle par les praticiens du marketing. Dans le cadre supervisé, selon l'objectif qu'on cherche à atteindre, il existe des critères consensuels à optimiser (AUC, p-value...), mais dans un cadre non supervisé, comme celui du partitionnement, le choix du nombre de classes, le choix de la méthode, voire de la quantité d'information retenue passent souvent par une approche exploratoire. De plus les méthodes géométriques reposant sur des calculs de distance, la gestion des variables qualitatives pose de nouvelles questions. Une stratégie fréquemment rencontrée consiste à discréteriser les variables numériques, puis à réaliser une analyse des correspondances multiples (ACM) pour calculer les distances euclidiennes à partir des axes factoriels quantitatifs.

1 Choix du nombre d'axes factoriels

La nécessité de choisir un nombre limité de plans factoriels pour les représentation induit l'idée discutable qu'il serait obligatoire ou préférable de ne conserver que quelques axes factoriels pour le calcul des distances. La notion de distance du chi deux, (équivalente à la distance euclidienne classique sur tous les axes de l'ACM), n'est pas toujours connue. L'interprétation difficile de la notion d'inertie, ou les transformations de valeurs propres censées redresser l'inertie expliquée par les premiers axes (transformation dite de Benzecri [0], par exemple) exigent des explications excessivement techniques pour les praticiens [1].

2 Choix du nombre d'axes et indice de Rand

L'indice de Rand est une mesure de proximité entre deux classifications [2]. Il repose sur le nombre d'individus pour lesquels les deux partitions sont "d'accord", soit classés ensemble dans les deux partitions, soit classés séparément dans les deux partitions. L'idée principale de l'interface proposée est de réaliser des partitions en conservant un nombre d'axes croissant et d'explorer graphiquement leurs proximités et si une stabilisation intervient à partir d'un certain nombre d'axes retenus.

3 L'alternative des modèles de mélange

Les modèles de mélanges de distributions discrètes sont des concurrents sérieux des méthodes strictement géométriques. L'implémentation d'une méthode de modèle de mélanges à partir du package MixAll [3] permet de la comparer à l'approche géométrique en évitant une démonstration trop théorique pour des praticiens.

4 L'interactivité pour favoriser l'exploration

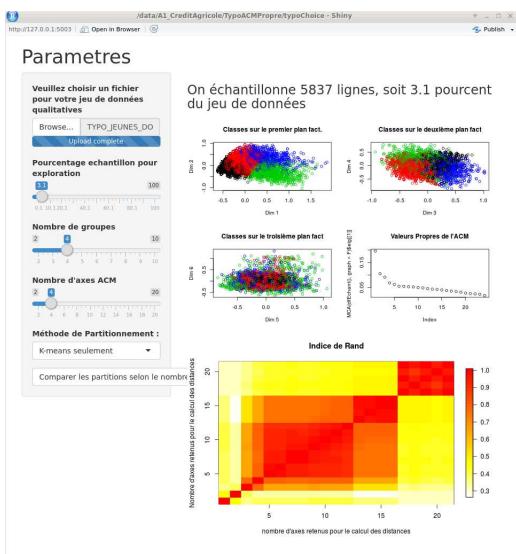


Figure 1: Choix interactif de typologies

Une interface R-Shiny a été développée. Elle permet de :

- Faire varier la taille de l'échantillon. Plus petit pour explorer, plus grand pour confirmer et exporter
- Faire varier le nombre d'axes de l'ACM retenus pour la réalisation de la typologie
- Représenter les classes sur les axes factoriels de son choix
- Représenter les proximités entre partitions suivant le nombre d'axes retenus
- Choisir une méthode basée sur les K-means et une autre mixant des K-means et de la CAH ou encore une méthode de modèle de mélange.

Après choix de la typologie, l'application permet :

- De représenter pour chacune des classes la répartition des modalités des variables les plus liées à la variable classe obtenue.
- D'exporter les données initiales augmentée des classes de la partition retenue.

Conclusion

L'interactivité permet à l'utilisateur d'explorer le nombre d'axes, de classes, la méthode employée, et de visualiser l'effet de ses choix sur la partition obtenue. Ce support ne remplace pas une compréhension profonde des méthodes mais permet d'y amener des praticiens pour qui l'exemple est souvent plus convaincant que la démonstration théorique.

Références

- [0] J.P. Benzecri. Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult. Les cahiers de l'analyse des données , 4 (3) 377-378, 1979.
- [1] S. Camiz and G. Coelho Gomes. Classification and Data Mining . Springer Berlin, Heidelberg, 2013. Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem.
- [2] G. Youness and G. Saporta. Une méthodologie pour la comparaison de partitions. Revue de statistiques appliquées , LII(1):97-120, 2004.
- [3] Serge Iovleff (2016). MixAll: Clustering using Mixture Models. R package version 1.2.0. <https://CRAN.R-project.org/package=MixAll>

Encapsuler une application R avec Docker

Vincent GUYADER

ThinkR
63 rue de la Pigaci  re
14000 Caen

vincent@thinkr.fr

Mots clefs : docker, R, application Shiny, package

Mod  le : Communication libre (20 minutes)

R  sum   :

Concevoir un outil R (un script, une application shiny...), c'est aussi se poser la question de sa p  rennit   dans le temps et de son d  ploiement actuel et futur. Comment s'assurer que ce qui a 茅t   con  u aujourd'hui sera encore utilisable dans 5 ans ? Comment faire en sorte que notre application puisse tourner sur tous les syst  mes d'exploitations actuels et futurs ?

Docker est un outil qui permet, entre autres, de r  pondre 脿 ces questions. Une application R et ses d  pendances sont encapsul  es dans un environnement d  di  , un container. Ce dernier peut 茅tre d  ploy   - et dupliqu   - ais  ment sur tous les syst  mes d'exploitation.

Nous vous pr  senterons aux travers de cas simples comment concevoir ce type de construction et comment d  ployer votre outil. Nous ouvrirons notre pr  sentation sur les autres usages possibles de Docker avec R.

**SI-APIMODEL : un système d'information développé avec R, Shiny, RPostgreSQL/PostGIS,
dplyr, DT, leaflet, ...**

A. Haddad^a, A. Kretzschmar^b

^a Unité de recherche BioSP
Centre de Recherches INRA-PACA
228 route de l'Aérodrome, Domaine Saint Paul - Site Agroparc , 84914 Avignon Cedex 9
abdelmalek.haddad@inra.fr

^b Unité de recherche BioSP
Centre de Recherches INRA-PACA
228 route de l'Aérodrome, Domaine Saint Paul - Site Agroparc , 84914 Avignon Cedex 9
andre.kretzschmar@inra.fr

Mots clefs : Apiculture, Santé des abeilles, Rucher, Performance, Base de données, Shiny Server.

Objectifs :

Lorsqu'on se propose de bâtir un système d'information «scientifique», avec en arrière-plan une base de données sous PostgreSQL/PostGIS et qu'on se propose de faire :

- la restitution sur le Web en utilisant les possibilités de Shiny (données et graphiques, calculs, visualisation des ruchers sur une carte, ...),
- la saisie des données, au travers du Web, par fichier ou par grille de saisie, avec un minimum de contrôles sur les champs,

alors la tâche de développement avec R et les packages candidats à la réalisation, semble a priori ardue. On est loin des outils «classiques» pour réaliser ce type de système. Si on ajoute les aspects «annexes» comme le choix d'un serveur Shiny (payant ou Open Source), les aspects sécurité et authentification (protection des échanges, filtrage des personnes pouvant opérer la saisie et/ou accéder à certaines données «sensibles»), alors il y a de quoi avoir des craintes quant à l'obtention de quelque chose d'exploitable sur le moyen terme.

Après une année de développement, nous avons abouti à la réalisation de SI-APIMODEL, un système concernant l'étude et le suivi de l'observation de miellées en rucher (lavandes, tournesol)

dans le cadre du projet APIMODEL, concernant principalement la santé et la performance des abeilles. On trouvera des informations sur celui-ci, en suivant le lien sur le site dédié, en passant par <https://w3.avignon.inra.fr/apimodel-apps/>, le site de SI-APIMODEL, ou encore en [1].

L'aspect architecture (logicielle et matérielle) a été privilégié dans ce document, ce qui donne une vue de l'organisation du «système» informatique comme une sorte de modèle autour duquel ont été développées diverses applications à base de code R. Ajouter d'autres applications basées sur le langage R est assez aisée.

Pour fixer les idées, nous avons choisi Shiny Open Source Edition pour déployer nos applications R/Shiny et rendre ces applications visibles sur le Web), car nos besoins sont amplement couverts par la version Open Source. Il existe différentes possibilités d'organisation et d'exécution des applications Shiny (le Web est riche d'exemples). Dans notre cas, SI-APIMODEL devait être accessible sur le Web pour une communauté de chercheurs, d'ingénieurs et de professionnels de l'apiculture. L'utilisation d'un serveur Web (Apache) nous a permis, au travers des fichiers de configuration, de réaliser une installation répondant à notre attente.

Le serveur Web HTTP Apache est utilisé comme (Reverse) «proxy» c'est-à-dire un relais : la requête HTTPS qui arrive sur Apache, à destination de Shiny Server, est «traduite» en une requête HTTP entre Apache et Shiny Server. Cette utilisation/traduction se fait au travers des fichiers de configuration d'Apache. A noter que dans notre architecture le logiciel Apache et le logiciel Shiny Server Open Source se trouvent sur la même machine, alors que le SGBDR PostgreSQL/PostGIS se trouve sur une autre machine (accessible seulement localement).

Perspectives :

Il reste à consolider l'utilisation PostgreSQL/PostGIS, dans les aspects de saisie (trigger et autres contraintes, gestion des erreurs), de connexion (pool, optimisation des accès).

La récupération des données par les utilisateurs, actuellement sous forme de fichiers csv, pourrait se faire à l'aide de Web Services. Pour cela QGIS semble un bon outil, SI-APIMODEL ayant un aspect SIG (les ruchers ont une composante géographique au travers de leurs coordonnées GPS).

Références :

- [1] Kretzschmar A., Maisonnasse A., Dussaubat C., Cousin M., Vidau C., 2016 Performances des colonies vues par les observatoires de ruchers Innovations Agronomiques 53, 81-93

The #SurfeR project : visualiser twitter avec R

Yan Holtz¹

¹ Freelance R & Dataviz, créateur du site *The R Graph Gallery*

6, Rue Jules Ferry, 34000 Montpellier

yan1166@hotmail.com

Mots clefs : R, TweetR, Dataviz

Résumé :

Ce « lightening talk » est un exemple d'analyse du réseau social tweeter réalisée avec R. Tous les tweets impliquant les hashtags **#surf**, **#kitesurf** et **#windsurf** ont été récupérés chaque jour ces ~200 derniers jours.

Les **~200 000 tweets** récupérés ont été filtrés, géo-localisés et caractérisés. Ils peuvent donc être utilisés afin de mieux comprendre les caractéristiques des amateurs de sport de glisse à travers le monde. Où habitent t'ils ? Ou et quand surfent t'ils ? Ou voyagent t'ils ? Que racontent t'ils ? Se connaissent t'ils ?

Tout en répondant à ces questions, l'exposé permettra de mettre en avant les fantastiques possibilités qu'offre R en terme de récupération et visualisation de l'information. Les principales slides présenteront :

- Tweeter : comment récupérer de l'info et comment surpasser les limitations de l'API ?
 - Récupération des infos complémentaires : localisation GPS, pays, continent, connexions.
 - Visualisation interactive de l'évolution du nombre de tweet.
 - Comment représenter géographiquement l'information.
 - Réalisation de nuages de mots afin de capter les messages principaux.
 - Réalisation de graphes en réseau afin d'appréhender l'organisation des comptes tweeter.

Avec une rencontre R à Anglet, impossible de repartir sans améliorer sa culture suRf !

Echantillon des visualisations proposées :

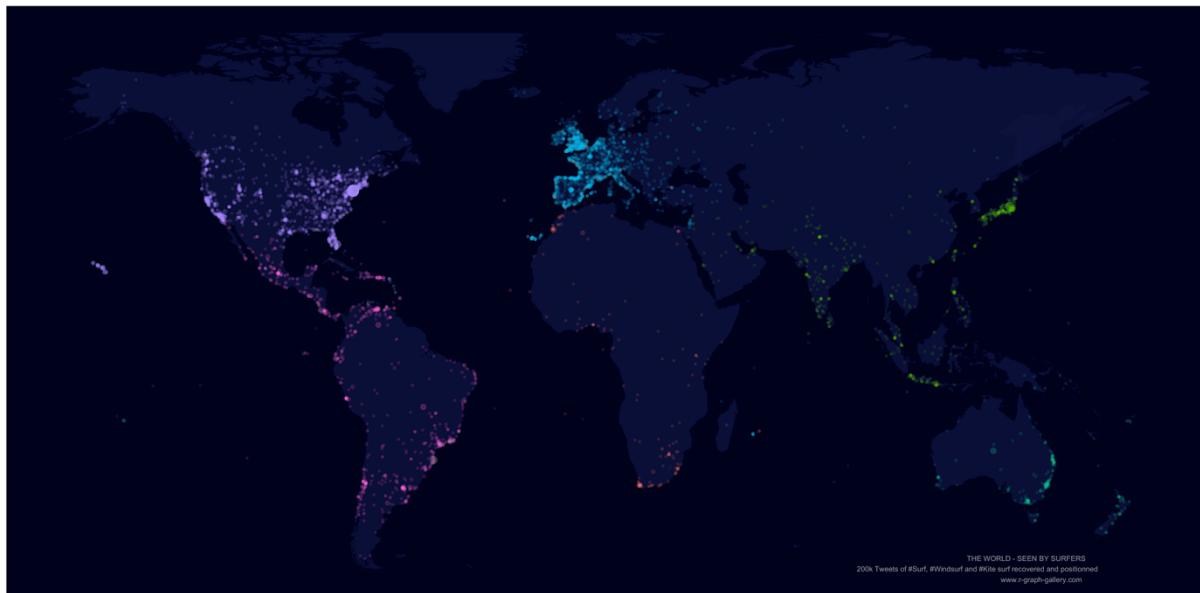


Figure 1: Localisation des tweets #Surf à travers le monde



Figure 2: Carte des voyageurs cherchant du #surf

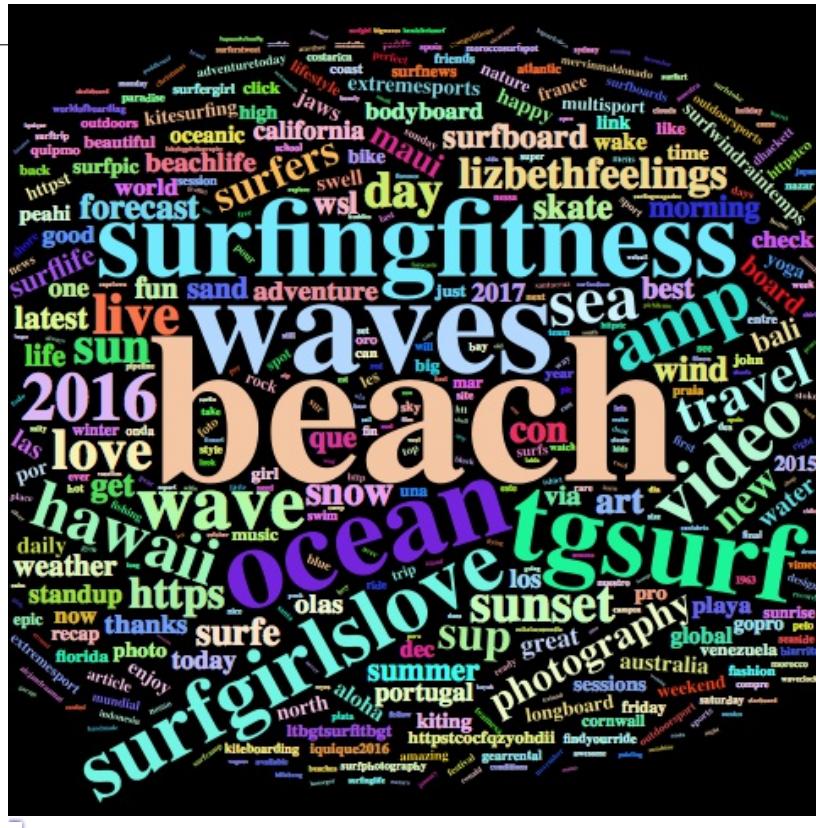


Figure 3: De quoi parlent les Surfeurs sur Tweeter?

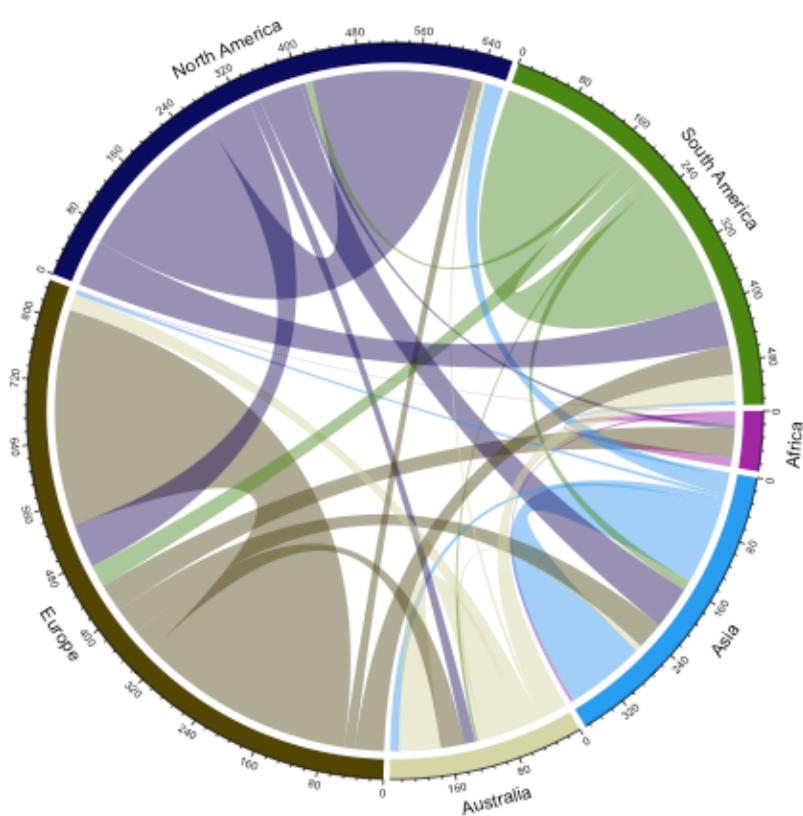


Figure 4: Bilan des déplacements des Surfeurs à travers le monde

Clouds, containers and R, towards a global hub for reproducible and collaborative open data science

K. Chine^a

^a RosettaHUB, Inc

karim.chine@rosettahub.com

Mots clefs: Cloud Computing, docker, collaboration, API, Python, Scala, reproducible research, education.

RosettaHUB aims at establishing a global open data science and open education meta cloud centered on usability, reproducibility, shareability and auditability. RosettaHUB platform and portal makes real-time collaboration ubiquitous and supports a wide range of open science-centric virtual social interactions.

RosettaHUB leverages public and private clouds and makes them easy to use for everyone. RosettaHUB's federation platform allows any higher education institution or research laboratory to create a virtual organization within the hub. The institution's members (researchers, educators, students) receive automatically active AWS accounts which are consolidated under one paying account, supervised in terms of budget and cloud resources usage, protected with safeguarding microservices and monitored/managed centrally by the institution's administrator. The cloud resources are generally paid for using the coupons provided by Amazon as part of the AWS Educate program. The Organization members' active AWS accounts are put under the control of a collaboration portal which simplifies dramatically everything related to the interaction with AWS and its collaborative use by communities of researchers, educators and students. The portal allows similar capabilities for Google Compute Engine, Azure, OpenStack-based and OpenNebula-based clouds.

RosettaHUB leverages Docker and allows users to work with containers seamlessly. Those containers are portable. When coupled with RosettaHUB's open APIs, they break the silos between clouds and avoid vendor lock-in. Simple web interfaces allow users to create those containers, connect them to data storages, snapshot them, share snapshots with collaborators and migrate them from one cloud to another. The RosettaHUB perspectives make it possible to use the containers to serve securely noVNC, RStudio, Jupyter and to enable those tools for real-time collaboration. Zeppelin and Spark-notebook and Shiny Apps are also supported. The RosettaHUB real-time collaborative containerized workbench is a universal IDE for data scientists. It makes it possible to interact in a stateful manner with hybrid kernels gluing together in a single process R, Python, Scala, SQL clients, Java, Matlab, Mathematica, etc. and allowing those different environments to share their workspace and their variables in memory. The RosettaHUB kernels and objects model break the silos between data science environments and make it possible to use them simultaneously in a very effective and flexible manner. A simplified reactive programming framework makes it possible to create reactive data science microservices and interactive web applications based on multi-language macros and visual widgets. A scientific web based spreadsheet makes it possible to interact with R/Python/Scala capabilities from within cells which includes variables import/export and variables mirroring to cells as well as the automatic mapping of any function in those environments to formulas invokable in cells. Spreadsheet cells can also contain code and code execution results making it become a flexible multi-language notebook. Ubiquitous docker containers coupled with the RosettaHUB workbench checkpointing capability and the logging

to embedded databases of all the interactions the users have with their environments make everything created within RosettaHUB reproducible and auditable.

The RosettaHUB's APIs (700+ functions) cover the full spectrum of programmatic interaction between users and clouds, containers and R/Python/Scala kernels. Clients for the APIs are available as an R package, a Python module, a Java library, an Excel add-in and a Word Add-in. Based on those APIs, RosettaHUB provides a CloudFormation-like service which makes it easy to create and manage as templates, collections of related Cloud resources, container images, R/Python/Scala scripts, macros and visual widgets alongside with optional cloud credentials. Those templates are cloud agnostic and they make it possible for anyone to easily create and distribute complex data science applications and services. The user with whom the template is shared can with one-click trigger the reconstruction and wiring on the fly of all the artifacts and dependencies. The RosettaHUB templates constitute a powerful sharing mechanism for RosettaHUB's e-Science and e-learning environments snapshots as well as for Jupyter/Zeppelin notebooks, shiny Apps, etc. RosettaHUB's marketplace transform those templates into products that can be shared or sold.

The presentation will be an overview of RosettaHUB and will discuss the results of the RosettaHUB/AWS Educate initiative which involved 30 higher education institutions and research labs counting over 3000 researchers, educators and students.

Références

- [1] www.rosettahub.com .
- [2] <http://bit.ly/rosettahub> .

Etude des règles d'assemblages des communautés : présentation d'une nouvelle approche méthodologique

Gaëlle Legras^a, Nicolas Loiseau^a, Jean-Claude Gaertner^a, Jean-Christophe Poggiale^b, Dino Ienco^c, Nabila Mazouni^a et Bastien Mérigot^d

^aUMR 241 EIO (UPF, IRD, Ifremer, IRD) B.P. 6570 - 98702 Faa'a - Tahiti - Polynésie française

^bAix Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography (MIO) UM 110 - 13288 Marseille - France

^cIRSTEA Montpellier, UMR TETIS - F-34093 Montpellier - France

^dUniversité de Montpellier, UMR MARBEC (CNRS, Ifremer, IRD, UM), 34203 Sète Cedex - France

Mots-clés : Règles d'assemblage des communautés, réseau fonctionnel, réseau de co-occurrence, indice de diversité, modularité

L'étude des règles d'assemblages qui structurent les écosystèmes est au centre de nombreuses recherches au sein de la communauté scientifique [1]. D'un point de vue méthodologique, de nombreuses études ont cherché à étudier cette question sur la base de l'étude des patterns de diversité fonctionnelle (*e.g.* [2], [3] et [4]). Cependant, les développements méthodologiques qui leur sont associés (*i.e.* dendrogramme, mesure du "convex hull volume" ou de l'hypervolume) possèdent de nombreuses lacunes, pouvant alors provoquer certaines erreurs d'interprétations dans les résultats obtenus ([5], [6]). Dans ce contexte, nous avons développé une nouvelle approche méthodologique basée sur la comparaison de deux réseaux complémentaires (réseau de co-occurrence des espèces et réseau fonctionnel) définis selon le calcul de leurs modularités respectives (calculées via l'algorithme d'optimisation de Louvain [7]). La similarité de ces deux réseaux (calculée via un indice de diversité modulaire, adapté de [8]) est comparée à celles obtenues sous modèles nuls afin de permettre aux utilisateurs d'identifier les principales règles d'assemblages structurant les communautés écologiques étudiées.

Basé sur un exemple empirique (identification des règles d'assemblages structurant des communautés d'abeilles au sein de trois fermes basées sur des stratégies de gestion agricole différentes), nous présenterons l'implémentation complète de notre méthode sous le logiciel R. En particulier, nous présenterons la fonction rendant capable le calcul de l'indice de diversité modulaire ainsi que sa comparaison avec ceux calculés via des modèles nuls (la construction de ceux-ci étant elle-aussi incluse dans la fonction). Enfin, nous présenterons comment nous pouvons utiliser le package *igraph* pour représenter graphiquement les réseaux obtenus et ainsi, permettre une visualisation plus intuitive des résultats qui en découlent (voir Figure 1 présentée ci-dessous).

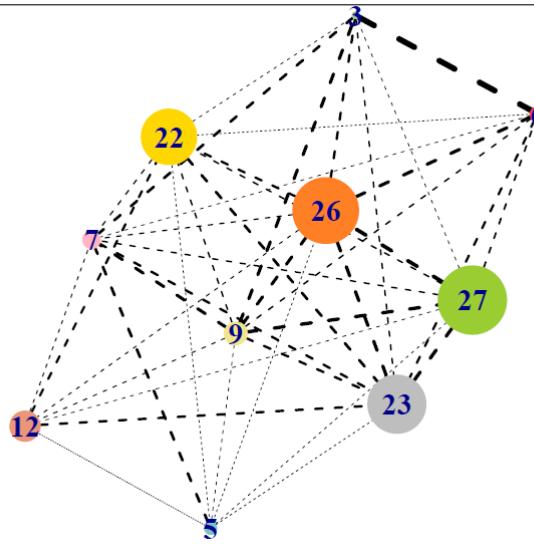


FIGURE 1 – Exemple de réseau pouvant être obtenu via la méthode présentée ci-dessus. L'épaisseur des liens est proportionnelle au degré de proximité entre chaque groupe d'espèces et le nombre dans chaque cercle représente le nombre d'espèce au sein de chaque groupe.

Références

- [1] Norman WH Mason, Cédric Lanoiselée, David Mouillot, Pascal Irz, and Christine Argillier. Functional characters combined with null models reveal inconsistency in mechanisms of species turnover in lacustrine fish communities. *Oecologia*, 153(2) :441–452, 2007.
- [2] Owen L Petchey and Kevin J Gaston. Functional diversity (fd), species richness and community composition. *Ecology Letters*, 5(3) :402–411, 2002.
- [3] William K Cornwell, Dylan W Schwilk, and David D Ackerly. A trait-based test for habitat filtering : convex hull volume. *Ecology*, 87(6) :1465–1471, 2006.
- [4] Benjamin Blonder, Christine Lamanna, Cyrille Violle, and Brian J Enquist. The n-dimensional hypervolume. *Global Ecology and Biogeography*, 23(5) :595–609, 2014.
- [5] J Podani. Convex hulls, habitat filtering, and functional diversity : mathematical elegance versus ecological interpretability. *Community Ecology*, 10(2) :244–250, 2009.
- [6] Nicolas Loiseau, Gaëlle Legras, Jean-Claude Gaertner, Philippe Verley, Pascale Chabanet, and Bastien Mérigot. Performance of partitioning functional beta-diversity indices : Influence of functional representation and partitioning methods. *Global Ecology and Biogeography*, 2017.
- [7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008, 2008.
- [8] Benoit Gauzens, Elisa Thébault, Gérard Lacroix, and Stéphane Legendre. Trophic groups and modules : two levels of group detection in food webs. *Journal of The Royal Society Interface*, 12(106) :20141176, 2015.

naniar: Data structures and functions for consistent exploration of missing data

A. Nicholas Tierney, B. Dianne Cook, C. Miles McBain

A Department of Econometrics and Business Statistics

A-Monash University

A-E762A Menzies Building, 20 Chancellors Walk, Monash University, Clayton, VIC 3800

nicholas.tierney@gmail.com

b Department of Econometrics and Business Statistics

B-Monash University

B-E762A Menzies Building, 20 Chancellors Walk, Monash University, Clayton, VIC 3800

dicook@monash.edu

c Bayesian Research and Applications Group

C-Queensland University of Technology

C-2 Level 8, Y block, George st, Brisbane, 4000, QLD

miles.mcbain@gmail.com

Mots clefs : Missing Data, Exploratory Data analysis, Imputation, Data Visualization, Data Mining, Statistical Graphics

Missing values are ubiquitous in data and need to be carefully explored and handled in the initial stages of analysis to avoid bias. However, exploring why and how values are missing is typically an inefficient process. For example, visualising data with missing values in ggplot2 results in omission of missing values with a warning, and base R silently omits missing values [1]. Additionally, imputed missing data are not typically distinguished in visualisation and data summaries. Tidy data structures described in [2] provide an efficient, easy and consistent approach to performing data manipulation and wrangling, where each row is an observation and each column is a variable. There are currently no guidelines for representing missing data structures in a tidy format, nor simple approaches to visualising missing values. This paper describes an R package, naniar, for exploring missing values in data with minimal deviation from the common workflows of ggplot and tidy data. Naniar builds data structures and functions that ensure missing values are handled effectively for plotting and summarising data with missing values, and examining the effects of imputation.

Références

- [1] Wickham, Hadley. 2009. Ggplot2: Elegant Graphics for Data Analysis. Springer.
- [2] Wickham, Hadley. 2014. “Tidy Data.” Journal of Statistical Software 59 (10).

Ch. Paroissin^{aet} G. Isaac^b

^a Université de Pau et des Pays de l'Adour
UFR Sciences et Techniques de Pau
64000 Pau
cparoiss@univ-pau.fr

^b Université de Pau et des Pays de l'Adour
Direction du Numérique
Pôle ARTICE
64000 Pau
guillaume.isaac@univ-pau.fr

Mots clefs : formation en ligne, formation hybride, modularité

1. Contexte

Les données et leurs traitements sont de plus en plus présentes dans tous les domaines et secteurs d'activités économiques et sociales. La maîtrise des méthodes statistiques et des outils informatiques associés est donc d'une haute importance, notamment en terme d'employabilité pour les étudiants. Le logiciel libre R apparaît dans ce contexte comme une réponse destinée à un large public. Il y a donc un réel besoin de formation à ce logiciel, que ce soit pour un public de spécialistes (étudiants en mathématiques appliquées, et plus spécialement en statistique) ou non, en formation initiale ou continue.

L'apprentissage d'un langage de programmation ou d'un logiciel peut être rébarbatif lorsqu'il est dispensé de manière classique en présentiel. Pour ce type de contenu pédagogique, une formation à distance en autonomie peut être tout à fait pertinente. Répondant à un appel à projets (interne à l'UPPA) en FOAD, Christian Paroissin a franchi le pas : pour le cours de R dispensé en 1ère année de la spécialité « Méthodes Stochastiques et Informatiques pour la Décision » du master « Mathématiques et Applications », il a choisi de passer du présentiel au distanciel. La version en présentiel se prêtait bien à cette transformation car l'ancienne version du cours était déjà structurée en séquences pédagogiques bien définies. La transposition du cours sur la plateforme moodle a donc été rapide. Le soutien du pôle ARTICE de l'UPPA et, en particulier, l'aide de Guillaume Isaac, ont été décisifs pour assurer la bonne mise en place du nouveau format.

Des rencontres avec d'autres collègues de l'UPPA ont fait émerger des besoins de formation à R pour des étudiants issus de deux autres masters, le master « Géographie, Aménagement, Environnement et Développement » (GAED) et le master « Sciences et Techniques des Activités Physiques et Sportives » (STAPS). Dès lors, le projet a pris une autre dimension et son développement a même franchi les Pyrénées.

2. Naissance d'une initiative pédagogique innovante

Progressivement, le projet initial s'est transformé et a débouché sur un dispositif pédagogique de plus grande envergure, en perpétuelle évolution. En effet, comme expliqué plus haut,

désormais le public visé n'est plus aussi homogène qu'au début : étudiants en master de mathématiques, de géographie et de sports. Comment concilier des publics aussi variés au sein d'une formation autour du même logiciel ?

Le dispositif repose sur un postulat de base : cette formation au logiciel R vient en complément de cours de statistique suivis par les étudiants durant leurs études. Les concepts statistiques sont donc déjà connus (à des degrés variés) et la formation vise à mettre en œuvre des méthodes statistiques avec R. Cependant, le contenu doit être différencié et adapté aux connaissances et aux besoins de chaque public. Ainsi, certaines parties s'adressent-elles spécifiquement à certains publics. La brique de base est donc un module portant sur une fonctionnalité précise de R. La flexibilité et la modularité de ce dispositif d'enseignement à distance permet de viser un large public. Actuellement, 25 modules ont été développés et permettent de répondre aux besoins des trois formations actuellement engagées dans le dispositif. En accord avec les responsables de formation, une sélection de 12 modules a été effectuée pour chaque formation. Déjà mis en œuvre pour les étudiants de mathématiques, le dispositif sera opérationnel à la rentrée 2017 pour les étudiants des deux autres masters.

3. Analyse pédagogique du dispositif

Chaque module est constitué de différentes ressources, avec au minimum des notes de cours et un test en ligne. Progressivement, tous les modules sont enrichis par d'autres contenus comme, par exemple, des tutoriels vidéos (ou capsules vidéos), l'objectif étant de mobiliser les différentes intelligences de l'apprenant. Le test est toujours la dernière étape d'un module et ne sert pas d'évaluation : il s'agit d'un test formatif. Cependant, un score minimum (75 % de la note maximale) est requis pour pouvoir accéder au module suivant (le nombre de tentatives est illimité). On se place ainsi dans une logique de progression par objectif. La réussite à un test de fin de module aboutit à l'obtention d'un badge. L'évaluation finale est réalisée à l'aide d'un test (formatif) dont la base de questions reprend l'ensemble des bases de questions pour les différents modules.

Afin de donner un rythme commun à tous les étudiants du même master, il est recommandé d'imposer une cadence hebdomadaire. Ainsi, un apprenant accède chaque semaine à un nouveau module sous réserve de la validation du module précédent et doit achever l'activité sur la semaine. Comme la formation est décomposée en 12 modules, la durée totale de la formation est de 13 semaines (correspondant à la durée d'un semestre à l'UPPA), la première semaine étant consacrée à la présentation du fonctionnement de la formation. En effet, afin d'animer la formation, trois séances en présentiel sont prévues (début, milieu et fin de formation). Elles sont complétées par un forum de discussion en ligne permettant un dialogue entre les apprenants d'une part et avec l'enseignant d'autre part.

4. (Encore) un MOOC ?

Quand on parle de ce projet à des collègues, cette question revient souvent. Non, ce n'est pas un MOOC. Certes, il s'agit d'un cours en ligne. Mais cette formation ne se veut ni massive, ni ouverte, pour des raisons structurelles d'organisation, de mise en œuvre pratique. En effet, comme expliqué plus haut, le contenu de la formation est adapté au mieux aux besoins de l'apprenant. Dans le cadre d'un étudiant de master, ces besoins sont définis par le responsable du diplôme. Au printemps 2017, cette formation vient d'être proposée pour la première fois aux doctorants de l'École Doctorale Sciences Exactes et Applications (ED SEA) de l'UPPA. Le caractère distanciel du dispositif permet de répondre à certaines spécificités d'un

établissement tel que l'UPPA : une université pluridisciplinaire et multi-site (5 campus éloignés les uns des autres). Ainsi, la formation est suivie par des doctorants dont les laboratoires sont à Pau, mais aussi à Mont de Marsan et à Anglet. Au début de la formation, les doctorants sont invités à exprimer leurs besoins qui peuvent être très différents les uns des autres. Une phase de diagnostic est donc mise en place en amont de la formation afin de co-construire le plan de formation. Une fois la liste des modules sélectionnés, les apprenants sont répartis dans des groupes : dans la mesure du possible, on essaye de proposer plusieurs modules en commun et de constituer différents groupes (avec le même plan de formation par groupe) afin de créer une synergie et un esprit collaboratif entre apprenants. Ils peuvent s'exprimer et s'entraider à travers un forum commun à tous les groupes (mais propre à chaque session de formation). Un des objectifs du dispositif, au niveau des doctorants, est aussi la constitution d'une communauté des utilisateurs de R à l'UPPA. Il est prévu un cadencement plus souple pour ce public : chaque semaine, un nouveau module est accessible sous réserve de valider le module précédent, mais ils ne sont pas contraints de l'achever dans la semaine.

5. Développement au transfrontalier

Suite à un accueil favorable de la part de collègues espagnols, les contenus d'une douzaine de modules ont été traduits en espagnol, grâce au soutien du programme Pyren (IDEFI), afin de proposer la formation aux étudiants des universités transfrontalières de l'UPPA : Université du Pays Basque (UPV/EHU), Université Publique de Navarre (UPNA) et Université de Saragosse (UZ). Cette version réduite de la formation sert de base aux dialogues nécessaires pour lever les verrous administratifs liés à cette situation atypique : le dispositif est hébergé à l'UPPA (moodle) et des accords bilatéraux sont en cours de négociation pour définir les conditions d'utilisation, d'animation et de validation de la formation pour les étudiants espagnols.

Pour l'année 2017-2018, une session de formation à R est prévue pour les doctorants des écoles doctorales de l'UZ et de l'UPNA. Il est également envisagé de proposer la formation aux étudiants du master de mathématiques commun à plusieurs universités du nord-ouest de l'Espagne, dont les trois universités transfrontalières de l'UPPA.

Un des challenges à venir est celui de la concrétisation de l'utilisation de la formation au transfrontalier. Si le succès est au rendez-vous, l'ensemble des modules feront donc l'objet d'une traduction afin de répondre aux différents besoins.

6. Perspectives et challenge à venir

Il reste encore bien d'autres publics à toucher. Il va de soi que les étudiants dans d'autres masters pourraient être aussi intéressés. On peut citer, comme exemple, le master « Chimie et Sciences du Vivant ».

En interne, il est également prévu de proposer la formation au personnel de l'UPPA (dans le cadre de la formation continue du personnel, cela se fera donc en lien avec la Direction des Ressources Humaines) sur le même principe que pour les doctorants.

Enfin, il est envisagé de proposer la formation à des personnes extérieures à l'UPPA dans le cadre du service de la Formation Continue (FORCO). Cette ouverture à l'extérieur pourra assurer la pérennisation du dispositif grâce aux frais d'inscriptions de ce dernier public.

Ce dispositif de formation modulaire est amené à rejoindre un autre projet porté également par Ch. Paroissin : il s'agit de la création d'un Centre de Consultation Statistique. Sur le même principe que d'autres structures existantes au sein de la Direction du Numérique, ce Centre aura pour vocation de fournir un appui méthodologique en statistique au personnel de l'UPPA dans leurs travaux de recherche. Le dispositif d'apprentissage à R correspondra donc au volet auto-formation de ce Centre. Des discussions sont actuellement en cours sur la création de Centre.

7. Portabilité du dispositif

Comme exemple de portabilité, on peut citer le langage de programmation Python. Ce langage est utilisé par différentes communautés comme les spécialistes d'analyse numérique (numériciens) ou les spécialistes de machine learning (data scientists). La base de Python constituerait alors le socle de connaissances communs à tous usagers du langage, puis différents autres aspects de Python pourraient former des modules plus dédiées à des usages spécifiques. On retrouve donc le principe de mutualisation de certains modules et de polylinéarité du dispositif.

Les outils informatiques peuvent donc naturellement faire l'objet d'un dispositif de formation tel que celui présenté ici. Cependant, cela ne constitue un champ exclusif de savoir. On pourrait très bien envisager un dispositif similaire pour d'autres domaines comme la biologie, la chimie, mais aussi l'économie, le droit, etc.

En effet, de manière générale, la conception de ce projet peut facilement être adaptée et transposée à d'autres formations dont les caractéristiques sont à peu près équivalentes : un corpus de savoirs découpé en modules qui puissent être recomposés pour aboutir à un plan de formation répondant aux besoins de l'apprenant (les besoins étant soit définis par un tiers – responsable de master, par exemple – ou par l'apprenant).

En résumé : en trois mots, où est l'innovation ?

- **modularité** : une formation à la carte, selon les connaissances et les besoins de l'apprenant ;
- **transdisciplinarité** : un public varié et motivé par un même objectif, l'acquisition de compétences en R ;
- **transfrontalier** : un dispositif déployé au niveau de l'UPPA et prochainement dans les universités transfrontalières partenaires de l'UPPA.

C3CO - Un package pour l'inference de la clonalité des cellules cancéreuses à partir du nombre de copies d'ADN

Morgane Pierre-Jean^a, Julien Chiquet^b, Henrik Bengtsson^c and Pierre Neuvial^d

^a Laboratoire de Mathématiques et de Modélisation d'Evry
Université d'Evry
23 boulevard de France, Evry
morgane.pierrejean@genopole.cnrs.fr

^c Epidemiology and Biostatistics
UCSF School of Medicine
550 16th. Street San Francisco
Henrik.Bengtsson@ucsf.edu

^b MIA Paris
UMR 518 AgroParistech/Inra
16 rue Claude Bernard, Paris
julien.chiquet@inra.fr

^d Institut Mathématiques de Toulouse
CNRS
118 Route de Narbonne, Toulouse
pierre.neuvial@math.cnrs.fr

Mots clefs : Bioinformatique, biostatistique , factorisation matricielle, cancer, clonalité, nombre de copies d'ADN

Contexte et modèle statistique

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.

Le package `c3co` implémente un problème de dictionary learning pour retrouver les sous-clones de tumeurs à partir de plusieurs profils de nombre de copies d'ADN issus d'expérience de puces à ADN ou de séquençage. Le modèle `c3co` peut être vu comme une extension du modèle `FLLat` de [1], et du modèle `e-FLLat` de [2] déjà implémentés.

Le package `c3co` est disponible depuis mars 2017 sur [github](https://github.com/pneuvial/c3co)¹.

Soit $y_{i\bullet}^1 \in \mathbb{R}^J$ et $y_{i\bullet}^2 \in \mathbb{R}^J$ les profils du nombre de copies parental (mineur et majeur) observés calculés à partir du nombre total de copies et du ratio allélique issus de puces SNP ou de séquençage:

$$y_{i\bullet}^m = \sum_{k=1}^p w_{ik} z_{k\bullet}^m + \epsilon^m, \text{ avec } m = 1, 2$$

- Les profils latents z^1 et z^2 sont supposés communs entre les profils observés
- Les poids w représentent la proportion de chacun des sous-clones dans les échantillons

On minimise:

$$\sum_{m=1,2} \sum_{i=1}^n \|y_{i\bullet}^m - \sum_{k=1}^p w_{ik} z_{k\bullet}^m\|^2 + \lambda_m \sum_{k=1}^p \sum_{s=1}^{S-1} |z_{k,s+1}^m - z_{k,s}^m|$$

sous la contrainte que $\forall i \sum_{k=1}^p w_{ik} = 1$ et $w_{ik} \geq 0$

Utilisation du package

Le package `c3co` permet de créer des profils synthétiques (profils latents) et générer une matrice de poids, résoudre le problème d'optimisation décrit ci-dessus et visualiser la matrice de poids inférée ainsi que les profils latents.

¹<https://github.com/pneuvial/c3co>

Création des profils synthétiques

A partir des données annotées du package `acnr`², la fonction `buildSubclones` peut être utilisée pour générer les sous-clones par rééchantillonnage. Les caractéristiques des sous-clones sont : la longueur des profils, le nombre de sous-clones, les points ruptures et les états des régions (gain, perte, normal).

```
dataAnnotTP <- acnr::loadCnRegionData(dataSet="GSE13372_HCC1143", tumorFraction=1)
dataAnnotN <- acnr::loadCnRegionData(dataSet="GSE13372_HCC1143", tumorFraction=0)
datSubClone <- buildSubclones(len=5000, dataAnnotTP, dataAnnotN, nbClones=3,
  bkps=list(c(100,250)*10, c(150,400)*10,c(150,400)*10),
  regions=list(c("(0,1)", "(0,2)","(1,2)",
    c("(1,1)", "(0,1)","(1,1)",
    c("(0,2)", "(0,1)","(1,1)"))))
```

On peut ensuite générer la matrice W et créer les profils hétérogènes.

```
W = getWeightMatrix(70, 30, nb.arch=3, nb.samp=10)
datList <- mixSubclones(subClones=datSubClone, W)
```

Inférence des paramètres du modèle

La méthode `c3co` a été appliquée au jeu de données créé ci-dessus. On choisit d'abord une grille de λ_1 and λ_2 et une grille de 2 à 5 pour le nombre de sous-clones.

```
lambda.grid <- seq(from=1e-6, to=1e-5, length=3)
nb.arch <- 2:5
parameters.grid <- list(lambda1=lambda.grid, lambda2=lambda.grid, nb.arch=nb.arch)
res <- c3co(datList, parameters.grid, verbose=FALSE)
```

Pour chaque p , `c3co` retient la meilleure combinaison (λ_1, λ_2) qui minimise le Bayesian Information Criterion (BIC) du modèle. La prochaine étape est de choisir p (le nombre de sous-clones). Suivant [1], nous avons calculé le pourcentage de variation expliquée (PVE) pour chacun des modèles, le p sélectionné est le dernier avant le plateau final de la courbe de la PVE en fonction du nombre de profils latents. Dans cet exemple le $p = 3$.

Figures

Pour des raisons de place nous ne montrons pas les figures produites mais il est possible de tracer la courbe de la PVE afin de choisir le bon nombre de profils latents, la matrice de poids ainsi que les profils latents.

```
pvePlot(res@fit, ylim=c(0.80,1))

best <- 2
res.clustTRUE = hclust(dist(cbind(W, 100-rowSums(W))),method="ward.D")
col = grDevices::colorRampPalette(RColorBrewer::brewer.pal(9, 'GnBu'))(100)
Wplot(res, idxBest = best, cexCol=0.2, main="PSCN", cexMain=0.5)
df <- createZdf(res, chromosomes=1, idxBest = best)
Zplot(df)
```

Références

- [1] Nowak, G., et al. (2011). A fused lasso latent feature model for analyzing multi-sample aCGH data. Biostatistics, 12(4), 776-791.
- [2] Masecchia, S., et al. (2013). A dictionary learning based method for aCGH segmentation. In ESANN.

²<https://github.com/mpierrejean/acnr>

The R package **bigstatsr**: Memory- and Computation-Efficient Statistical Tools for Big Matrices

F. Privé (a), H. Aschard (b) and M.G.B. Blum (a)

(a) Laboratoire TIMC-IMAG – Université Grenoble Alpes - CNRS – Faculté de Médecine - 38706 La Tronche cedex - France

`florian.prive.21@gmail.com` – `florian.prive@univ-grenoble-alpes.fr` – `michael.blum@univ-grenoble-alpes.fr`

(b) Département de Génomes et Génétique – Centre de Bioinformatique, Biostatistique et Biologie Intégrative – Institut Pasteur - 25-28 Rue du Dr Roux, 75015 Paris - France

`hugues.aschard@pasteur.fr`

Mots clefs : Statistics, Big Data, Memory-mapping, Parallelism.

Abstract

The *R* package **bigstatsr** (<https://github.com/privetf/bigstatsr>) provides functions for fast statistical analysis of large-scale data encoded as matrices. The package can handle matrices that are too large to fit in memory. The package **bigstatsr** is based on the format **big.matrix** provided by the *R* package **bigmemory** (Kane, Emerson, and Weston 2013).

The package **bigstatsr** enables users with laptop to perform statistical analysis of several dozens of gigabytes of data. The package is fast and efficient because of four different reasons. First, **bigstatsr** is memory-efficient because it uses only small chunks of data at a time. Second, special care has been taken to implement effective algorithms. Third, **big.matrix** objects use memory-mapping, which provides efficient accesses to matrices. Finally, as matrices are stored on-disk, many processes can easily access them in parallel.

The main features currently available in **bigstatsr** are:

- singular value decomposition (SVD) and randomized partial SVD (Lehoucq and Sorensen 1996),
- sparse linear and logistic regressions (Zeng and Breheny 2017),
- sparse linear Support Vector Machines,
- column-wise linear and logistic regressions tests,
- matrix operations,
- parallelization / apply.

References

Kane, Michael J, John W Emerson, and Stephen Weston. 2013. “Scalable Strategies for Computing with Massive Data.” *Journal of Statistical Software* 55 (14): 1–19. doi:10.18637/jss.v055.i14.

Lehoucq, Rich Bruno, and D. C. Sorensen. 1996. “Deflation Techniques for an Implicitly Restarted Arnoldi Iteration.” *SIAM Journal on Matrix Analysis and Applications* 17 (4). Society for Industrial; Applied Mathematics: 789–821. doi:10.1137/S0895479895281484.

Zeng, Yaohui, and Patrick Breheny. 2017. “The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R,” January. <http://arxiv.org/abs/1701.05936>.

Visualisation interactive d'arbres de décision avec visNetwork

T. Robert^a and B. Thieurmel^b

^{ab}Datastorm

60 rue Etienne Dolet - 92240 Malakoff

^atitouan.robert@datastorm.fr

^bbenoit.thieurmel@datastorm.fr

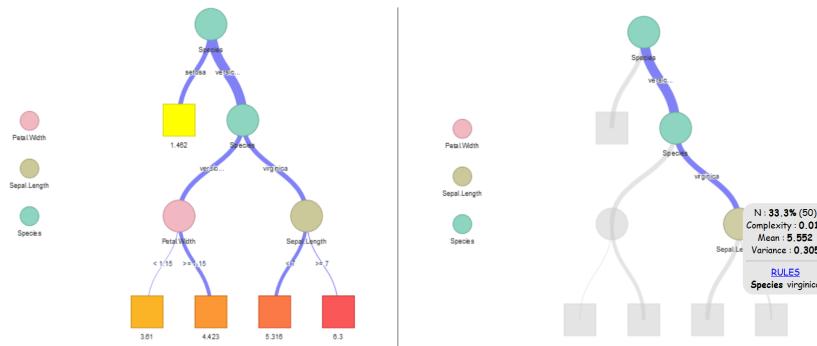
Mots clefs : Réseaux, Visualisation, Javascript, Shiny, Arbre, Classification

Lancé en 2015 à l'occasion des rencontres R de Grenoble, le package **visNetwork** a depuis trouvé son public dans la communauté. Basé sur la librairie javascript **vis.js**[1], il permet de créer, visualiser et intéragir avec des réseaux.

En plus des fonctionnalités initiales (réseaux orientés / hiérarchiques, stylisation, paramétrisation de la physique), de nombreuses améliorations ont été implémentées depuis :

- Mise en surbrillance des noeuds et agrégation dynamique(*visOptions*)
- Interface avec le package **igraph**[2] (*visIgraph* et *visIgraphLayout*)
- Implémentation de méthodes pour modifier le réseaux dans **shiny**[3] (*visNetworkProxy*)

De plus, nous finalisons actuellement la fonction *visTree* permettant d'explorer dynamiquement un arbre de régression ou de classification résultant du package **rpart**[4].



Le package est disponible sur le **CRAN** ainsi que sur **github** à l'adresse suivante : <https://github.com/datastorm-open/visNetwork>.

Références

- [1] Almende B.V. Javascript library vis.js. <http://visjs.org>.
- [2] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. <http://igraph.org>
- [3] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.0. <https://CRAN.R-project.org/package=shiny>
- [4] Terry Therneau, Beth Atkinson and Brian Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. <https://CRAN.R-project.org/package=rpart>

Outil d'interprétation de score

Elena SALETTE

Datastorm

60, rue Etienne Dolet

92240 Malakoff

elena.salette@datastorm.fr

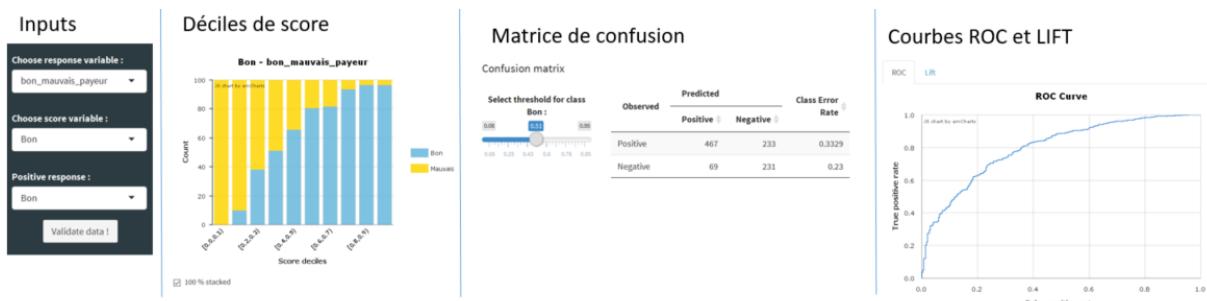
Mots clefs : Statistique, Scoring.

Les méthodes usuelles de scoring telles que les forêts aléatoires ou les réseaux de neurones par exemple mènent à des résultats difficilement interprétables. En effet, le fait de ne pas obtenir de coefficients associés aux variables du modèle ou dont l'interprétation n'est pas immédiate comme pour la régression logistique ne rend pas aisée la compréhension des résultats. Par ailleurs, la détermination de la valeur seuil du score déterminant l'appartenance ou non à une classe n'est généralement pas simple et dépend de la finalité du modèle.

Nous avons donc développé un outil d'aide à l'interprétation d'un score, se traduisant en une application **shiny**. En partant d'un jeu de données contenant la variable réponse, des variables explicatives (prises en compte ou non dans la modélisation initiale) ainsi que le score obtenu, cet outil permet de répondre notamment aux questions suivantes :

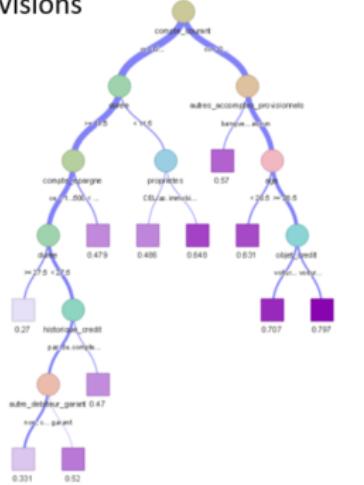
- Quelles sont les variables qui ont le plus contribué à l'élaboration du score ?
- Quelle est la répartition du score et sa performance selon les valeurs d'une des variables explicatives ?
- Quel choix faire pour la valeur du seuil de score ?
- Quelles variables ont le plus d'influence sur le score ?
- Peut-on expliquer ou catégoriser les bonnes ou les mauvaises prédictions ?

Mesure de la performance

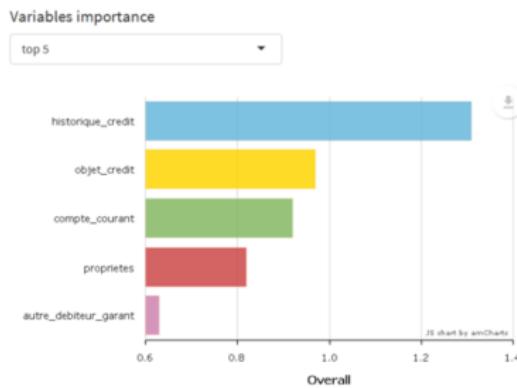


Interpretation

Arbre CART expliquant la variable réponse, le score ou la qualité des prévisions



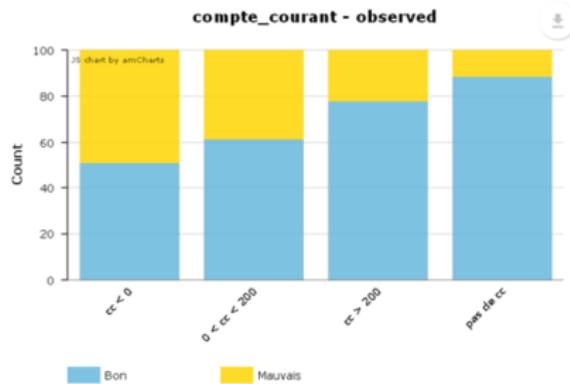
Importance des variables



Matrice de confusion partielle en fonction des modalités ou des intervalles d'une variable explicative

	Modality	N	Accuracy	Real Pos. (%)	Of which pred. Pos. (%)	Real Neg. (%)	Of which pred. Neg. (%)
1	cc < 0	274	61.31%	50.73%	27.34%	49.27%	96.30%
2	0 < cc < 200	269	57.62%	60.97%	40.24%	39.03%	84.76%
3	cc > 200	63	52.38%	77.78%	65.31%	22.22%	7.14%
4	pas de cc	394	50.00%	88.32%	93.10%	11.68%	23.91%

Répartition de la réponse en fonction d'une variable explicative



Références

- [1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2016). shiny: Web Application Framework for R. R package version 0.14.2. <https://CRAN.R-project.org/package=shiny>

Comment je suis devenue crolute

Maëlle Salmon^a, Scott A. Chamberlain^b

a ISGlobal

ISGlobal, Doctor Aiguader 88, E-08003 Barcelona, Spain

maelle.salmon@yahoo.se

b rOpenSci

rOpenSci, University of California, Berkeley CA, 94720, USAUniversity of California, Berkeley CA, 94720, USA

scott@ropensci.org

Mots clefs : API, requêtes http, paquet, R6.

ropenaq est une interface R pour l'API d'OpenAQ qui permet l'accès à des données libres de qualité de l'air. Le paquet permet l'utilisation des différents services offerts par l'API: obtenir des mesures de concentrations de polluants comme l'ozone à un endroit donné et obtenir des listes de pays/villes/stations disponibles pour un paramètre donné de qualité de l'air. Le paquet **ropenaq** fournit les données en tant que **data.frames**, les rendant compatibles avec de nombreux outils comme le **tidyverse**. Le paquet, développé sur Github, a été accepté par le projet rOpenSci après une revue par les pairs et est sur CRAN. Bien que le paquet avait atteint un état stable, nous avons décidé de remplacer une de ses dépendances-clés, le client HTTP **httr** développé à RStudio, par **cruk**, un autre client HTTP développé à rOpenSci. Pourquoi ? Est-ce parce que Crul est le nom d'une planète dans les films Star Wars?

Notre première motivation était la vitesse. Toutes les données sur les PM2.5 (particules de taille inférieure à 2.5 micromètres) aux Etats-Unis pendant une semaine représentent plus de 80 000 mesures. L'API elle-même a un système de pagination ne permettant que le retour de 10 000 mesures par consultation, donc une telle requête implique au moins 8 consultations de l'API. Le paquet **cruk** permet des requêtes HTTP asynchrones, ce qui n'est pas possible avec **httr**, donc utiliser **cruk** a permis un gain de temps substantiel. Dans une petite expérience comparative réalisée avec le paquet **microbenchmark**, nous avons fait 100 répétitions de la même demande de données par **ropenaq** avec ou sans requêtes asynchrones, l'API d'OpenAQ permettant jusqu'à 10 requêtes en même temps. La demande testée était d'obtenir les 20 538 mesures pour Delhi en 14 jours, demandant 1000 valeurs par page avant de ne pas surcharger l'API de demandes. La médiane du temps a été réduit par un facteur d'environ 6 de 37 à 6 secondes. Une telle augmentation de vitesse peut être cruciale, par exemple si le paquet **ropenaq** était utilisé dans un **shinydashboard** interactif, ou même pour simplement obtenir des réponses plus rapides dans la console.

Notre seconde motivation, ou bien plutôt un avantage collatéral du changement est le fait que **cruk** soit construit autour de classes **R6**. À cause de cela, beaucoup d'étapes des requêtes http et du traitement des réponses sont réalisées avec des méthodes d'une classe **R6**, par exemple la classe **HttpRequest**, au lieu d'utiliser des fonctions. Cela signifie qu'en développant un paquet tel **ropenaq** il suffit d'importer des classes, pas une longue liste de fonctions. En plus de cet argument, les classes **R6** ont une documentation et une logique qui sont aisément compréhensibles, non seulement pour les développeurs habitués à la programmation orientée object, mais aussi pour les développeurs moins expérimentés qui peuvent apprécier le fait que les méthodes soient documentées dans la documentation de la classe. En outre, après avoir créé un objet de **cruk** il est possible d'écrire “\$” et la fonction d'auto-complétion de RStudio permet de voir les méthodes disponibles pour cet objet.

Comme illustration, voici un extrait de code demandant et traitant du contenu d'une URL, et extrayant le status de la demande, à la fois avec **httr** et **cruk**. Le code **cruk** a une ligne de plus mais seulement un import alors que le code **httr** en a trois.

```
library("crul")
library("httr")
# URL
url <- "https://httpbin.org/get"

# code httr
response <- httr::GET(url)
parsed_content <- httr::content(response, as = "text")
status <- httr::status_code(response)

# code crul
client <- crul::HttpClient$new(url = url)
res_get <- client$get()
parsed_content <- res_get$parse()
status <- res_get$status_code
```

Dans cet oral, nous présenterons d'abord succinctement les projets rOpenSci et OpenAQ, et expliquerons ensuite les avantages du changement de `httr` à `crul`. Ce sera l'occasion de présenter la syntaxe du paquet `crul`, et d'introduire les requêtes HTTP asynchrones. Nous conclurons la présentation par des visualisations réalisées avec des données téléchargées via le paquet `ropenaq`. A la fin de cette présentation, l'auditoire saura pourquoi je suis devenue une habitante de Crul, et comment faire de même.

Linkspotter : outil interactif d'exploration et de visualisation de corrélations

Alassane Samba

Orange Labs

2 avenue Pierre Marzin, 22300 Lannion, France

alassane.samba@orange.com

Mots clefs : Corrélation, Information mutuelle, Graphe dynamique, Visualisation, Interface utilisateur.

L'exploration des relations entre les variables est une étape importante dans le processus d'exploration de données. Il permet au Data scientist de mieux comprendre les phénomènes décrits dans les jeux de données. Cela aide également à éviter le sur-apprentissage et à se prémunir contre le « concept drift » dans la modélisation.

Afin de faciliter l'exploration des données, Linkspotter est un package R offrant plusieurs fonctionnalités permettant d'analyser et de visualiser de manière exhaustive en utilisant un graphe toutes les corrélations bi-variées d'un fichier de données.

Ses fonctionnalités principales sont:

- le calcul de plusieurs matrices de corrélation correspondant à différents coefficients
- le partitionnement des variables par apprentissage non supervisé.

Il offre également une interface utilisateur complète, conviviale et personnalisable permettant de :

- visualiser les corrélations à l'aide d'un graphe (les variables correspondant aux noeuds et les corrélations correspondant aux arcs). C'est une nouvelle approche de visualisation qui est proposée plus conviviale que l'affichage d'une matrice de corrélations coloriée.
- visualiser la distribution de chaque variable grâce à son histogramme ou à son diagramme en barres.
- visualiser un lien entre un couple de variables à l'aide de nuage de points, de boîtes-à-moustaches, etc.

En outre, un nouveau coefficient de corrélation que nous appelons Maximal Normalized Mutual Information (MaxNMI) basé sur une discrétisation supervisée Equal-Freq des variables continues est également introduit par Linkspotter. L'intérêt de ce coefficient est qu'il peut être calculé et comparé quel que soit le type de couple de variables (continue vs catégorielle, continue vs continue, catégorielle vs catégorielle).

Tout en offrant de nouveaux algorithmes et méthodes, Linkspotter utilise et combine harmonieusement des fonctionnalités provenant d'autres packages R, à savoir infotheo, minerva, energy, mclust, shiny, visNetwork et rAmCharts.

Le code [1] et une démonstration [2] de Linkspotter sont disponibles en ligne.

Références

[1] <https://github.com/sambaala/linkspotter>

[2] <http://linkspotter.sigmant.net>

M. Saumard^a

^aConservatoire National des Arts et Métiers
Laboratoire CEDRIC, Equipe MSDMA
292 Rue Saint-Martin, 75003 Paris
matthieu.saumard@lecnam.net

Mots clefs : Big Data, Clustering, MongoDB et R, données fonctionnelles.

Financements: Projet e-FRAN: Territoire Calculant en Bourgogne- Franche-Comté.

1 Introduction

Le jeu vidéo Mathador est un jeu de calcul mental du type le compte est bon. Dans une épreuve classique, on a 5 nombres de lancer et un nombre cible compris entre 1 et 99. Le but de l'épreuve est d'obtenir le nombre cible en utilisant les nombres du lancer et les quatre opérations courantes (addition, soustraction, multiplication, division). Le score associé de l'épreuve tient compte des opérations utilisées, à savoir l'utilisation d'une division rapporte 3 points, une soustraction 2, une addition ou une multiplication 1 point. Le coup Mathador consiste à utiliser les 4 opérations en 4 lignes de calcul.

2 Les données

Nous recevons un dump chaque Lundi de la totalité de la base de données MongoDB conçues pour ce jeu.

2.1 Importation

Pour importer les données, nous utilisons le package R mongolite créé par Jeroen Ooms (2014) [3] . Nous utilisons ensuite la fonctionnalité iterate pour importer les données dans R. Nous utilisons cette fonction car les documents sont emboités et ne permettent pas directement de créer des data frame.

2.2 Fouille de données

Il existe différents jeux : solo et chrono. Les données sont enregistrées différemment suivant le jeu et l'application utilisé : web, apps android, ios. Les calculs effectués par le joueur sont enregistrés sauf les calculs effacés par la touche effacer. Plusieurs variables sont enregistrées : les nombres du lancer, la cible, les calculs, les dates de jeu, abandon, réponse correct, entre autres. On prévoit d'ici la fin de l'année scolaire 2017 plus de 700 000 calculs effectués par les classes e-FRAN étudiées.

3 Construction de profils

Un des but est de construire des profils de calculant. On définit des indicateurs à partir des calculs, par exemple, les pourcentages d'utilisation des différentes opérations. Puis on applique des algorithmes de clustering. Une approche par données fonctionnelles est aussi considérée. On construit alors des courbes d'évolution et on les groupe en utilisant par exemple [1] et [2] .

Références

- [1] Francesca Ieva, Anna M Paganoni, Davide Pigoli, and Valeria Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401–418, 2013.
- [2] Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics ; Data Analysis*, 71:92–106, 2014.
- [3] Jeroen Ooms. The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv preprint arXiv:1403.2805*, 2014.

A. Siberchicot^a, B. Spataro^a et S. Delmotte^a

^aLaboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558

Univ Lyon, Université Lyon 1, CNRS

F-69622 Villeurbanne, France

{aurelie.siberchicot, bruno.spataro, stephane.delmotte}@univ-lyon1.fr

Mots clefs : serveur *shiny*, recherche reproductible

Aujourd’hui, de plus en plus de journaux scientifiques demandent à ceux qu’ils publient de rendre publics et accessibles librement les outils et données relatifs à leur article. Dans certains champs disciplinaires, les chercheurs sont déjà habitués à diffuser leurs données et leurs sources, alors que, dans d’autres, il est encore nécessaire de les convaincre à partager ces éléments avec leur communauté. Des solutions en ligne ou locales, gratuites ou pas, fleurissent pour proposer des sites, des plateformes, des serveurs pour rendre accessible ce qui constitue la matière première des articles scientifiques et garantir la reproductibilité des résultats.

Dans cette logique de recherche reproductible, une réflexion pour une solution informatique a été engagée au sein du laboratoire de Biométrie et Biologie Évolutive (LBBE) de Lyon, pour proposer à l’ensemble de ses chercheurs (plus d’une centaine de personnes travaillant sur des problématiques scientifiques allant de l’écologie à la génomique en passant par la biométrie) un espace pour diffuser les données et les sources de leur recherche. Inscrit dans un contexte académique, le service développé devait être conforme aux politiques de sécurité des systèmes d’informations des établissement d’enseignement supérieur et de recherche. Ces contraintes peuvent se traduire par la localisation physique des données et des services numériques dans les locaux du LBBE, et non chez un prestataire extérieur. Une solution gratuite devait également être préférée.

Grâce aux ressources matérielles disponibles et fort de ses compétences variées, le pôle informatique du LBBE a déployé, pour répondre à cette problématique, un serveur *shiny* local, dans sa version libre et gratuite. Techniquelement, ce serveur est installé sur une machine virtuelle et déroule un scénario de réinstallation quotidiennement en suivant une recette *puppet*¹. Les chercheurs du LBBE (et uniquement eux) peuvent ainsi diffuser aussi bien des applications *shiny* que des documents en *R markdown* via ce serveur *shiny*. Pratiquement, chaque élément diffusé fait l’objet d’un dépôt *Subversion*², hébergé sur un hôte différent et mis à jour quotidiennement.

Encore jeune, cet outil est en constante évolution : les différentes étapes d’installation du serveur et de sa mise à jour feront progressivement l’objet d’une automatisation. Mais l’expérience informatique menée en interne par le pôle informatique du LBBE constitue déjà un exemple solide de mise en place d’un serveur *shiny* local. Il propose aux chercheurs du LBBE une alternative au développement assez systématique de packages *CRAN*. En effet, l’accroissement important du nombre de packages *R* ces dernières années (environ 6000 en mars 2015, environ 8000 en

¹*puppet* est un logiciel de gestion de la configuration de serveur.

²*Subversion* (*svn* en abrégé) est un logiciel de gestion de versions.

mars 2016, plus de 10000 aujourd’hui), amène à se poser des questions sur la pertinence de nouveaux packages dans le cadre de la diffusion d’un outil ou d’un résultat scientifique. Enfin, le développement local de ce service permet d’être au plus près des contraintes et demandes liées aux spécificités des recherches menées au sein du LBBE.

Le serveur *shiny* du LBBE est accessible à l’adresse <http://lbbe-shiny.univ-lyon1.fr/>.

MareyMap Online : une application shiny pour estimer les taux de recombinaison grâce à l'approche par carte de Marey.

A. Siberchicot^a, A. Bessy^b, L. Guéguen^a et G. Marais^a

^aLaboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558
Univ Lyon, Université Lyon 1, CNRS
F-69622 Villeurbanne, France
{aurelie.siberchicot, laurent.gueguen, gabriel.marais}@univ-lyon1.fr

^badrienbessy@hotmail.com

Mots clefs : taux de recombinaison méiotique, carte de Marey, application *shiny*

La recombinaison méiotique est un processus largement étudié en biologie. De nombreuses études décrivent ce processus et attestent que les taux de recombinaison peuvent largement varier le long du génome. Une bonne description de ces taux de recombinaison est donc importante pour décrire cette étape de la méiose mais aussi pour comprendre ses nombreuses implications sur l'évolution du génome.

Plusieurs approches méthodologiques ont été développées pour estimer les taux de recombinaison méiotique. La plus populaire est l'approche dite par carte de Marey. Cette méthode repose sur la comparaison de la carte génétique (sur laquelle les distances entre gènes correspondent à des fréquences de crossing-over, en centimorgan) et de la carte physique (sur laquelle les distances entre gènes sont en nombre de nucléotides, le plus souvent en mégabase) d'un chromosome donné, pour une espèce donnée. Sur le graphe des positions génétiques en fonction des positions physiques de tous les gènes d'un même chromosome, l'estimation locale des taux de recombinaison est alors donnée par la pente locale de la courbe ajustée sur les points de cette carte de Marey [1]. Cette méthode d'estimation des taux de recombinaison est d'autant plus plébiscitée qu'il est aujourd'hui facile d'obtenir ces cartes, grâce à la publication de séquences complètes du génome pour de nombreux eucaryotes.

Plusieurs outils ont implémenté l'approche par carte de Marey. Mais la plupart de ces outils sont centrés sur un seul organisme et ne permettent pas à un utilisateur d'analyser ses propres données. Le package *R MareyMap* [2] [3] [4] est le seul outil qui, d'une part, met à disposition les cartes de plusieurs organismes et, d'autre part, propose d'analyser de nouvelles cartes fournies par l'utilisateur. Ce package a été développé sous la forme d'une interface graphique en *tcl/tk* et de nombreuses fonctionnalités y sont disponibles pour paramétrier et affiner les modèles d'estimation des taux de recombinaison.

Suite au succès du package *MareyMap*, nous avons développé une application *shiny*, *MareyMap Online* [5], qui vient étendre les possibilités d'utilisation de l'approche par carte de Marey. *MareyMap Online* est une version simplifiée de *MareyMap* ; elle est accessible depuis un navigateur web (ne nécessite donc pas l'installation de *R* et de packages, juste d'une connexion) et accompagne l'utilisateur pas à pas pour le chargement et le nettoyage des données (Figure 1), l'estimation des taux de recombinaison et l'export des résultats. Trois méthodes sont proposées pour estimer les taux de recombinaison : les fenêtres glissantes (sliding windows), le

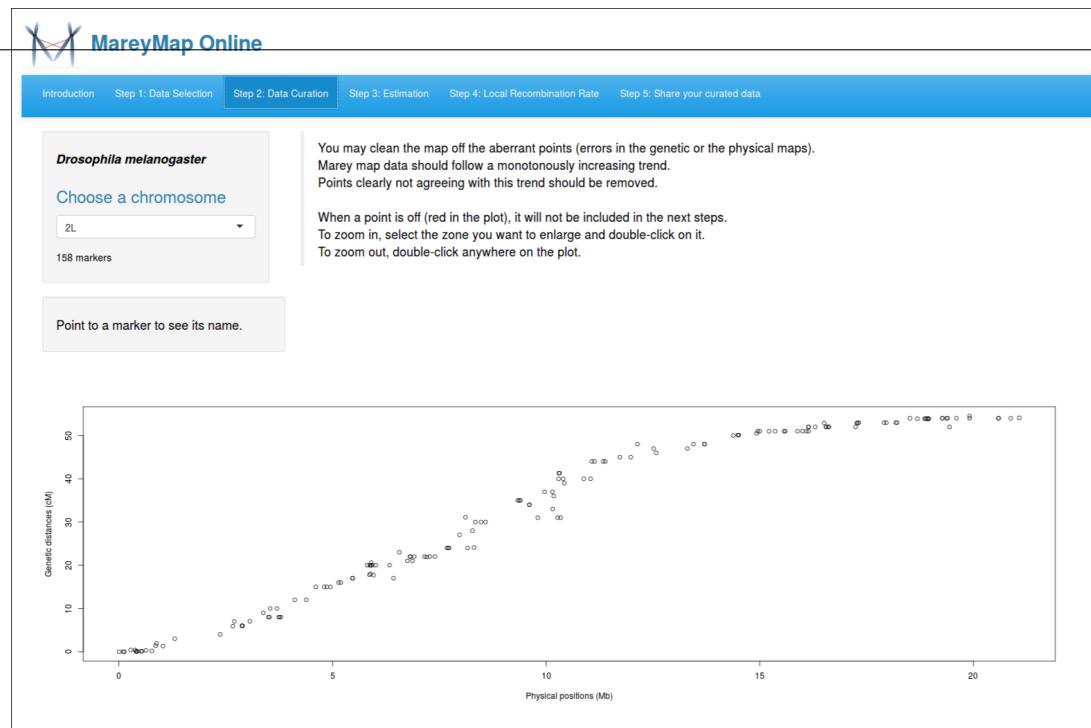


Figure 1: Onglet de l'étape du nettoyage des données dans *MareyMap Online*. Exemple de la carte de Marey du chromosome 2L de *Drosophila melanogaster*.

loess et les splines cubiques (cubic splines). L’automatisation du choix de certains paramètres de modélisation facilite l’exploration des données par des biologistes qui ne seraient pas familier des cartes de Marey. *MareyMap Online* est également enrichi de jeux de données concernant 8 organismes ce qui rend possible la réalisation d’une métá-analyse sur les taux de recombinaison méiotique. Enfin, les utilisateurs qui ont analysé leurs propres données sont encouragés à laisser leur adresse électronique, pour que leurs cartes soient par la suite intégrées à *MareyMap Online* et rendues disponibles à la communauté.

Cette application est disponible à l’adresse <http://lbbe-shiny.univ-lyon1.fr/MareyMapOnline/>.

Références

- [1] Chakravarti, A. (1991). A graphical representation of genetic and physical maps: the Marey map. *Genomics*, **11**, 219-222
- [2] Rezvoy, C., Charif, D., Guéguen, L., Marais, G. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, **23**, 2188-2189
- [3] Siberchicot, A., Rezvoy, C., Charif, D., Guéguen, L., Marais, G. (2015). The MareyMap package. Version 1.3.3. *CRAN*.
- [4] <https://lbbe.univ-lyon1.fr/-MareyMap-.html>
- [5] Siberchicot, A., Bessy, A., Guéguen, L., Marais, G. (2017). MareyMap Online: a user-friendly web service for estimating recombination rates using physical and genetic map. *Genome Biology and Evolution*, en cours de révision.

Travaux sous tension : traitement d'un volume important de données et outils de visualisation

S. Issad^a and B. Thieurmel^b

^aRTE

aDirection R&D-I - Projet Smartlab

9 rue de la Porte de Buc, BP 561 - 78005 VERSAILLES

samir.issad@ rte-france.com

^bDatastorm

60 rue Etienne Dolet - 92240 Malakoff

benoit.thieurmel@datastorm.fr

Mots clefs : hdf5, shiny, visualisation, javascript, séries temporelles

La réalisation de travaux sous tension (TST) est un atout essentiel pour RTE (gestionnaire du Réseau de Transport Electrique haute et très tension français) parce qu'elle permet de limiter les consignations et donc l'indisponibilité des ouvrages concernés. Dans ce contexte, RTE, en collaboration avec Datastorm, a développé une web-application **shiny**[1] destinée aux métiers de la maintenance, en charge des TST, afin de leurs donner des outils statistiques d'aide à la décision :

- Accès moderne, rapide, pratique, et ergonomique aux données et aux résultats
- Filtrage patrimoniale, temporelle et spatial
- Visualisation dynamique
- Exportation des résultats / génération de rapport

Partant d'un historique de transits sur 3 ans, disponible au pas de temps 5 minutes, et divisé en une centaine de fichiers plats pour un total de 60 Go environ, les données ont préalablement été restructurées et stockées au format **hdf5**[2] en utilisant le package **r hdf5**[3]. Ce système permettant, en plus d'un gain de place non négligeable, des performances très intéressantes pour requêter sur un sous-ensemble de données.

Ce projet a également permis d'ajouter de nouvelles fonctionnalités au package de visualisation **rAmCharts**[4], développé par les équipes de Datastorm :

- Synchronisation des séries temporelles
- Exportation côté client des graphiques javascript
- Module de visualisation de séries temporelles de grande dimension (communication entre **shiny** et la librairie javascript pour agréger plus ou moins finement les données en fonction du zoom utilisateur)



Le package **rAmCharts**[4] est disponible sur le **CRAN** ainsi que sur **github** à l'adresse suivante : <https://github.com/datastorm-open/rAmCharts>.

Références

- [1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.0. <https://CRAN.R-project.org/package=shiny>
- [2] The HDF Group, Hierarchical data format version 5, 2000-2010, <http://www.hdfgroup.org/HDF5>
- [3] Bernd Fischer and Gregoire Pau (2016). rhdf5: HDF5 interface to R. R package version 2.18.0.
- [4] Jeffery Petit, Antanas Marcelionis, Benoit Thieurmel, Elena Salette and Titouan Robert (2017). rAmCharts: JavaScript Charts API Tool. R package version 2.1.3. <https://CRAN.R-project.org/package=rAmCharts>

FactoInvestigate - Description automatique de résultats d'analyse factorielle

S. Thuleau^a and F. Husson^b

^aEulidia

20 rue Thérèse - 75001 Paris
simon.thuleau@gmail.com

^bAgrocampus Ouest

65 rue de St-Brieuc - 35042 Rennes
husson@agrocampus-ouest.fr

Mots clefs : Analyse factorielle, ACP, Analyse des correspondances, Description automatique.

FactoInvestigate est un package qui propose d'aider à l'interprétation des résultats d'analyse factorielle comme l'ACP, l'AFC ou l'ACM. Le package prend en entrée un objet résultat d'analyse factorielle obtenu par une des fonctions du package *FactoMineR* et fournit en sortie un document au format Word, PDF ou html contenant les principaux indicateurs, les graphes essentiels et une ébauche de commentaires des résultats. Ainsi, de façon automatique, une première interprétation des résultats est proposée.

La fonction *Investigate*, qui est la principale fonction du package, utilise de nombreux arguments pour choisir les critères, les graphes ou les indicateurs qui serviront à l'élaboration du commentaire. Ces arguments sont définis par défaut mais ils sont modifiables afin de répondre aux besoins de l'utilisateur.

La fonction *Investigate* fait appel à plusieurs sous-fonctions permettant l'automatisation de chaque étape de la description :

- détection d'individus aberrants ou extrêmes (outliers, décrits puis retirés de l'analyse),
- estimation du nombre de composantes à interpréter,
- illustration par des graphiques judicieux (coloration éventuelle des individus selon la variable qualitative pertinente),
- optimisation de la lecture des graphiques (affichage d'un nombre d'éléments adaptés afin d'éviter la surcharge visuelle),
- mise en évidence de groupes d'individus et caractérisation de ces groupes par les variables, etc.

Chacune de ces étapes est accompagné d'un commentaire synthétique et sur mesure.

Le fichier RMarkdown utilisé pour générer le fichier de sortie Word, PDF ou html peut également être conservé, afin d'ajuster à loisir les graphes ou commentaires. Il donne également l'intégralité du script R nécessaire à la reproduction des analyses.

En outre, *FactoInvestigate* détecte la langue de votre ordinateur et afin de délivrer un commentaire en français pour les utilisateurs francophones.

Voici à titre d'exemple un extrait des commentaires produits de façon automatique sur le jeu de données décathlon du package *FactoMineR*.

Analyse en Composantes Principales

Jeu de données decathlon

Ce jeu de données contient 41 individus et 13 variables, 2 variables quantitatives sont illustratives, 1 variable qualitative est illustrative.

1. Observation d'individus extrêmes

L'analyse des graphes ne révèle aucun individu singulier.

2. Distribution de l'inertie

L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.

Les 2 premiers axes de l'ACP expriment **50.09%** de l'inertie totale du jeu de données ; cela signifie que 50.09% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important, et le premier plan représente donc convenablement la variabilité contenue dans une grande partie du jeu de données actif. Cette valeur est supérieure à la valeur référence de **37.85%**, la variabilité expliquée par ce plan est donc significative (cette inertie de référence est le quantile 0.95 de la distribution des pourcentages d'inertie obtenus en simulant 1753 jeux de données aléatoires de dimensions comparables sur la base d'une distribution normale).

Du fait de ces observations, il serait tout de même probablement préférable de considérer également dans l'analyse les dimensions supérieures ou égales à la troisième.

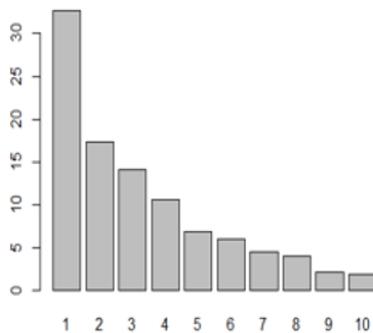


Figure 2 - Décomposition de l'inertie sur les composantes principales de l'ACP

Cet extrait est fourni par la fonction *Investigate* avec les lignes de code suivantes :

```
> library(FactoMineR)
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13)
> library(FactoInvestigate)
> Investigate(res.pca, document="word_document")
```

Références

- [1] Husson, F., Lê, S., Pagès, J. (2016). Analyse de données avec R, 2ème édition, *Presses Universitaires de Rennes (PUR)*.

Modélisation bayésienne d'une chronologie d'évènements archéologiques et analyse des chaines de Markov à l'aide du package 'ArchaeoPhases'

Marie-Anne Vibet and Anne Philippe

Laboratoire de Mathématiques Jean Leray

Université de Nantes

2, rue de la Houssinière 44000 Nantes

marie-anne.vibet@univ-nantes.fr

anne.philippe@univ-nantes.fr

Mots clefs : Statistique bayésienne, Archéologie, Chronologie d'évènements.

Les archéologues utilisent maintenant très largement la modélisation bayésienne afin de construire des chronologies d'évènements archéologiques. Cette modélisation consiste en l'estimation de dates à partir d'observations qui sont les mesures d'age réalisées sur les prélevements (datation par radiocarbone, archéomagnétisme...) et à partir de l'information *a priori* provenant du contexte archéologique, géologique ou tout autre considération.

L'inférence bayésienne est une méthode d'estimation probabiliste, l'information obtenue sur les paramètres estimés étant donnée par une densité de probabilité appelée la densité *a posteriori*. Généralement, la forme analytique d'une telle densité de probabilité n'est pas facile à exprimer. Cependant, il est possible de simuler une chaîne de Markov dont la loi stationnaire est la densité *a posteriori* désirée. En effet, l'algorithme de Monte Carlo par chaînes de Markov (MCMC) permet de simuler un échantillon de la densité *a posteriori* pour des modèles bayésiens de grandes dimensions [3].

Le package ArchaeoPhases, que nous avons développé, fournit une liste de fonctions permettant le traitement statistique de dates archéologiques ou de groupes de dates. Ces fonctions sont basées sur le traitement des chaînes de Markov dont la loi stationnaire est la densité *a posteriori* de la série de dates. De telles chaînes de Markov peuvent être générées par différents logiciels tels que 'ChronoModel' (voir <http://www.chronomodel.fr>), 'Oxcal' (voir <https://c14.arch.ox.ac.uk/oxcal.html>) ou 'BCal' (voir <http://bcal.shef.ac.uk/>).

Dans ce package, nous proposons de nouveaux outils statistiques permettant d'analyser des groupes de dates. A partir de ces dates, on peut estimer le début (respectivement la fin) du groupe par minimum (respectivement le maximum) des dates du groupe. Nous caractérisons un groupe par l'intervalle de recouvrement (time range interval), qui correspond au plus court intervalle contenant les dates du groupe. Pour des successions de groupes, on propose des estimations des périodes de transition et on teste la présence d'un gap (d'un hiatus) entre deux groupes de dates successifs. D'autre part on estime le rythme d'occurrence d'évènements archéologiques.

Dans cette présentation, nous montrerons comment la statistique bayésienne s'applique à la construction de chronologies d'évènements archéologiques. Nous considérerons le site archéologique de Canima Abajo, à Cuba, dans lequel deux épisodes d'enterrements ont été retrouvés l'un en dessous de l'autre et séparé par une strate contenant des coquillages [4]. Nous traiterons cet exemple afin d'illustrer les principales fonctions du package 'ArchaeoPhases'. Ainsi, nous verrons comment utiliser le package 'coda' pour vérifier la convergence des chaînes de Markov simulées et l'absence d'autocorrélation entre les valeurs successives des chaînes. Nous observerons le rythme d'occurrence des enterrements [1,2]. Enfin, nous estimerons les périodes d'activités liées aux deux cimetières [2] : estimation des intervalles de recouvrements, de l'intervalle de transition et test de l'existence d'un gap.

Nous présenterons également l'application shiny associée au package 'ArchaeoPhases'. Cette application web permet aux non-utilisateurs de R (la plupart des archéologues) de bénéficier des fonctions de ce package sans

avoir à manipuler R.

Références

- [1] Dye, T.S. (2016). Long-term rhythms in the development of Hawaiian social stratification. *Journal of Archaeological Science*, 71, 1–9
- [2] Philippe A. and Vibet M.-A.(2017). Analysis of Archaeological Phases using the CRAN Package ‘ArchaeoPhases’. HAL, hal-01347895, version 3
- [3] Robert C. and Casella (2009). Introducing Monte Carlo Methods with R. Springer Science & Business Media.
- [4] Roksandic M. et al. (2015) Radiocarbon and Stratigraphic Chronology of Canímar Abajo, Matanzas. Cuba. *Radiocarbon*, Vol 57, N 5, 755-763.

NOM	PRENOM	EMAIL
AIROLDI	Cyrille	olivier.clericy@developpement-durable.gouv.fr
BARON	Alexis	marie-michele.ladjyn@univ-paris1.fr
BARTOLO	François	bartolo@methodomics.com
BASCHET	louise	louise.bachet@capionis.com
BAUDROT	Virgile	virgile.baudrot@univ-lyon1.fr
BELDAME	Diane	Diane@thinkr.fr
BELLOUARD	Angéline	angeline.bellouard@yahoo.fr
BERTON	Sylvie	sylvie.berton@univ-pau.fr
BERTRAND	Julie	julie.bertrand@inserm.fr
BESSE	Sylvere	sylvere.besse@univ-pau.fr
BESSOU	Antoine	antoine.bessou@cnamts.fr
BITH	Tiphaine	Tiphaine.BITH@agencerecherche.fr
BOUCHET	Adelin	adelin.bouchet@cnamts.fr
BOUYSSIE	Marianne	Marianne.Bouyssie@keyrus.com
BRAZEILLES	Remi	remi.brazeilles@danone.com
BRENAC	Nathalie	Nathalie.Brenac@afpa.fr
BROC	Camilo	camilo.broc@univ-pau.fr
BRU	Noelle	noelle.bru@univ-pau.fr
CAILLAUD	Marie-anne	marie-anne.caillaud@capionis.com
CALLENS	Aurélien	aurelien.callens@orange.fr
CANNAVACCIUOLO	Mario	m.cannavacciuolo@groupe-esra.com
CASANOVA	Sandrine	sandrine.casanova@tse-fr.eu
CAUQUIL	Laurent	laurent.cauquil@inra.fr
CAZEAU	Géraldine	geraldine.cazeau@anses.fr
CHAMPION	Magali	marie-helene.gbaguidi@parisdescartes.fr
CHAVENT	Marie	marie.chavent@inria.fr
CHECCHI	Alix	alix.checchi@anses.fr
CHINE	Karim	karim.chine@rosettahub.com
COLONGO	David	david.colongo@hyphen-stat.com
COMBET	Romain	romain.combet@sanofi.com
COUALLIER	Vincent	vincent.couallier@u-bordeaux.fr
D'AMICO	Frank	frank.damico@univ-pau.fr
DAKPO	K Hervé	k-herve.dakpo@inra.fr
DALLEMANE	Luc	idallemagne@maltem.com
DAMITIO	Emma	contact@audap.org
DAZIN	Franck	fdazin@inpi.fr
DEHMAN	Alia	alia.dehman@hyphen-stat.com
DELAIGUE	Olivier	olivier.delaigue@irstea.fr
DEMOISY	Bruno	bruno.demoisy@univ-pau.fr
DESCHAMPS	Aline	aline.deschamps@dacta.fr
DESJEUX	Yann	yann.desjeux@inra.fr
DEVILLE	Marion	marion.deville@unige.ch
DHORNE	Thierry	thierry.dhorne@univ-ubs.fr
DOS SANTOS PEREIRA	Fabio	Fabio.Dos-santos-pereira@developpement-durable.gouv.fr
DRAY	Stéphane	stephane.dray@univ-lyon1.fr
DROUILHET	Rémy	remy.drouilhet@upmf-grenoble.fr
EBERT	Anthony	ac.ebert@qut.edu.au
EL ALAOUI FARIS	Moulay Driss	moulay-driss.elalaouifar@airliquide.fr
ETERRADOSSI	Olivier	olivier.etteradossi@mines-ales.fr
ETIENNE	Marie-Pierre	marie.etienne@agroparistech.fr
FAUGERE	Julien	jfaugere@rd.loreal.com
FAVIER	Clément	marie.lebris@mixscience.eu
FAY	colin	contact@colinfay.me
FELTIN	Clément	clement.feltin@rte-france.com
FRIGOT	Eric	eric.frigot@anses.fr

GABIEL	Julien	Gabriel.Julien@keyrus.com
GAGNAIRE	Thomas	tgagnaire@inbox.fr
GAMBOA	Bastien	bastien.gamboa@yahoo.fr
GENOLINI	Christophe	cg@rplusplus.com
GENUER	Robin	robin.gener@u-bordeaux.fr
GERDS	Thomas	tag@biostat.ku.dk
GOMBIN	Joel	joel.gombin@gmail.com
GOUDE	Yannig	yanniggoude@yahoo.fr
GORUDINE	Jean-Luc	Jean-Luc.Gourdine@inra.fr
GORRET	Damien	damien.gourretbaumgart@gmail.com
GOUSSEF	Matthieu	matthieu.gousseff@univ-ubs.fr
GRANGER	Victoria	victoria.granger@enedis.fr
GRISONI	Marie-Lise	mlgrisoni@hotmail.com
GUYADER	Arnaud	arnaud.guyader@upmc.fr
GUYADER	Vincent	vincent@thinkr.fr
HADDAD	Abdelmalek	abdelmalek.haddad@inra.fr
HASSEN-KHODJA	Cédric	cedric.hassen-khodja@mri.cnrs.fr
HOLTZ	Yan	yan1166@hotmail.com
HONTEBEYRIE	Sophie	sophie.hontebeyrie@univ-pau.fr
HUSSON	François	husson@agrocampus-ouest.fr
IMMER	Anika	anika.immer@qmail.com
JAUNATRE	Kevin	kevin.jaunatre@univ-ubs.fr
JEFFROY	Guillaume	guillaume.jeffroy@gmail.Com
JEGOU	Nicolas	nicolas.jegou@univ-rennes2.fr
JOURDAN	Astrid	aj@eisti.eu
JULIEN	Rémi	remi.julien@fr.imptob.com
KERMORVANT	Claire	claire.kermorvant@univ-pau.fr
KLAIC-LOPEZ	Rafael	rklaiclopez@inbox.fr
KOUDOU	Efoevi Angelo	Efoevi.Koudou@univ-lorraine.fr
KRETZSCHMAR	André	andre.kretzschar@inra.fr
LALANNE	Yann	yann.lalanne@univ-pau.fr
LAMBRECHTS	Hugues	hugues.lambrechts@gmail.com
LATOUCHE	Aurélien	aurelien.latouche@curie.fr
LAUGEROTTE	Alexandra	alexandra.laugerotte@ipsen.com
LAURENT	Julie	laurent@methodomics.com
LECONTE	Eve	eve.leconte@tse-fr.eu
LEGRAIS	Gaëlle	legras.gaelle@gmail.com
LEGRIS	Maxime	laurent.lenours@idele.fr
LEROI	Fanny	fanny.leroy@cnamts.fr
LIQUET	Benoit	benoit.liquet@univ-pau.fr
LUCBERNET	Robin	rlucbernet@maltem.com
MACH	Nuria	nuria.mach@inra.fr
MAIGNE	Elise	elise.maigne@inra.fr
MAPELLA	Clément	clement.mapella@lecnam.net
MARCHAND	Miranda	miranda.marchand@rte-france.com
MARCHIONNI	David	david.marchionni@sanofi.com
MARQUE	sebastien	sebastien.marque@capionis.com
MAUGER	Emmanuelle	emmanuelle.mauger@chanel-corp.com
MEDDIS	Alessandra	alessandra.meddis@gmail.com
MEZIERES	Sophie	sophie.mezieres@univ-lorraine.fr
MORENO	vincent	vincent.moreno@hupi.fr
MOROLDO	Marco	marco.moroldo@inra.fr
MORVAN	Aurèle	aurele@creativedata.fr
MOUGENOT	Flavien	flavien.mougenot@mixscience.eu
MOURER	Alex	marie-michele.ladjyn@univ-paris1.fr
MOURGIART	Bastien	bastien.mourguiart@etud.univ-pau.fr

NATHOO	Farouk	nathoo@uvic.ca
NICHOLAS	Tierney	nicholas.tierney@gmail.com
NOLAIN	Patrick	patrick.nolain@sanofi.com
NOT	Claire	mc.not@adil30.org
PAROISSIN	Christian	christian.paroissin@univ-pau.fr
PATOU	Alison	Alison.Patou@keyrus.com
PATOUILLE	Brigitte	karine.lecuona@u-bordeaux.fr
PAYET	Vincent	vpayet@isara.fr
PEAUCELLIER	Sophie	claude.couture@cesdip.fr
PIERRE-JEAN	Morgane	morgane.pierrejean@genopole.cnrs.fr
PILAUD	Thomas	thomas.pilaud@gmail.com
PINAIRE	Jessica	jessica.pinaire@chu-nimes.fr
PONTET	Célia	c.pontet@terresinovia.fr
POUEY	Jérôme	jerome.pouey@santepubliquefrance.fr
REBOURS	Pierre	prebours@maltem.com
REGNIER	Catherine	Catherine.Sager@davigel.fr
RIGAL	Francois	francois.rigal@univ-pau.fr
ROBERT	Titouan	titouan.robert@datastorm.fr
RONDEAU	Pacale	pacale.rondeau@danone.com
ROQUERE	romain	romain.roquefere@hupi.fr
ROUVIERE	Laurent	laurent.rouvriere@univ-rennes2.fr
SAINT-PIERRE	Philippe	Philippe.Saint-Pierre@math.univ-toulouse.fr
SALETTE	Elena	elena.salette@datastorm.fr
SALMON	Maëlle	maelle.salmon@yahoo.se
SAMBA	Alassane	alassane.samba@orange.com
SANTINI	Josie	josie.santini@ecofog.gf
SAUMARD	Matthieu	matthieu.saumard@gmail.com
SAUSSAC	Mathilde	mathilde.saussac@anses.fr
SAVY	Nicolas	marie.lebris@mixscience.eu
SEGALINI	Audrey	audrey.segalini@hyphen-stat.com
SIBERCHICOT	Aurélie	aurelie.siberchicot@univ-lyon1.fr
TAP	Julien	julien.tap@danone.com
TENTELIER	Cédric	cedric.tentelier@univ-pau.fr
THEBAULT	Anne	Anne.THEBAULT@anses.fr
THIEURMEL	Benoit	benoit.thieurmel@datastorm.fr
THULEAU	Simon	simon.thuleau@gmail.com
TU NGUYEN	Minh	minh-tu.nguyen@hupi.fr
VIALLON	Vivian	vivian.viallon@univ-lyon1.fr
VIBET	Marie-Anne	marie-anne.vibet@univ-nantes.fr
VOIRIN	Dominique	dominique.voirin@edf.fr