The R User Conference, useR! 2017
July 4-7 2017
Brussels, Belgium

Book of Contributed Abstracts

Contents

I Talks

II Lightning Talks

III Posters

Part I

Talks

MCMC Output Analysis Using R package **mcmcse**

Dootika Vats[1*], *James M. Flegal*[2], *John Hughes*[3], *Galin L. Jones*[1]

** Presenting author*
1. University of Minnesota, Twin-Cities
2. University of California, Riverside
3. University of Colorado, Denver

Keywords: MCMC, stopping rule, effective sample size.

Webpages: https://cran.r-project.org/web/packages/mcmcse/index.html, http://users.stat.umn.edu/~vatsx007/

Markov chain Monte Carlo (MCMC) is a method of producing a correlated sample in order to estimate expectations with respect to a target distribution. A fundamental question is when should sampling stop so that we have good estimates of the desired quantities? The key to answering these questions lies in assessing the Monte Carlo error through a multivariate Markov chain central limit theorem. This talk presents the R package **mcmcse**, which provides estimators for the asymptotic covariance matrix in the Markov chain CLT. In addition, the package calculates a multivariate effective sample size which can be rigorously used to terminate MCMC simulation. I will present the use of the R package **mcmcse** to conduct robust, valid, and theoretically just output analysis for Markov chain data.

difNLR: Detection of potentional gender/minority bias with extensions of logistic regression

Adéla Drabinová[1] and Patricia Martinkova[2]

1. *Faculty of Mathematics and Physics, Charles University, Prague*
2. *Institute of Computer Science, Czech Academy of Sciences, Prague*

Keywords: detection of item bias, differential item functioning, psychometrics, R

Webpages: https://CRAN.R-project.org/package=difNLR, https://CRAN.R-project.org/package=ShinyItemAnalysis, https://shiny.cs.cas.cz/ShinyItemAnalysis/

The R package difNLR has been developed for detection of potentially unfair items in educational and psychological testing, analysis of so called Differential Item Functioning (DIF), based on extensions of logistic regression model. For dichotomous data, six models have been implemented to offer wide range of proxies to Item Response Theory models. Parameters are obtained using non-linear least square estimation and DIF detection procedure is performed by either F or likelihood ratio test of submodel. For unscored data, analysis of Differential Distractor Functioning (DDF) based on multinomial regression model is offered to provide closer look at individual item options (distractors). Features and options are demonstrated on three data sets. The package is designed to correspond to difR package (one of the most used R libraries in DIF detection, see Magis, Béland, Tuerlinckx, & De Boeck (2010)) and currently is exploited by ShinyItemAnalysis (Martinková, Drabinová, Leder, & Houdek, 2017) which provides graphical interface offering detailed analysis of educational and psychological tests.

References

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Martinková, P., Drabinová, A., Leder, O., & Houdek, J. (2017). ShinyItemAnalysis: Test and item analysis via shiny. Retrieved from shiny.cs.cas.cz/ShinyItemAnalysis/; https://CRAN.R-project.org/package=ShinyItemAnalysis

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. CBE-Life Sciences Education, 16(2). https://doi.org/10.1187/cbe.16-10-0307

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., et al (2017). Development and validation of the homeostasis concept inventory. CBE-Life Sciences Education, 16(2). https://doi.org/10.1187/cbe.16-10-0305

useR! 2017 Oral Presentation

factorMerger: a set of tools to support results from post-hoc testing

Agnieszka Sitko[1] and Przemysław Biecek[1]

1. Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Keywords: analysis of variance (ANOVA), hierarchical clustering, likelihood ratio test (LRT), post-hoc testing

Webpages: https://github.com/geneticsMiNIng/FactorMerger

ANOVA-like statistical tests for differences among groups are available for almost a hundred years. But for large number of groups the results from commonly used post-hoc tests are often hard to in- terpret. To deal with this problem, the **factorMerger** package constructs and plots the hierarchical relation among compared groups. Such hierarchical structure is derived based on the Likelihood Ratio Test and is presented with the Merging Paths Plots created with the **ggplot2** package. The current implementation handles one-dimensional and multi-dimensional Gaussian models as well as binomial and survival models. This talk presents the theory and examples for a single-factor use cases.

Interactive graphs for blind and print disabled people

A. Jonathan R. Godfrey[1] *Paul Murrell*[2] *and Volker Sorge*[3]

1. *Institute of Fundamental Sciences, Massey University, New Zealand*
2. *Department of Statistics, University of Auckland, New Zealand*
3. *Department of Computer Science, University of Birmingham, United Kingdom*

Keywords: accessibility, exploration, interactivity

Descriptions of graphs using long text strings are difficult for blind people and others with print disabilities to process; they lack the interactivity necessary to understand the content and presentation of even the simplest statistical graphs. Until very recently, *R* has been the only statistical software that has any capacity for offering the print disabled community any hope of support with respect to accessing graphs. We have levered off the ability to create text descriptions of graphs and the ability to create interactive web content for chemical diagrams to offer a new user experience.

We will present the necessary tools that (1) produce the desired graph in the correct form of a scalable vector graphic (SVG) file, (2) create a supporting *XML* structure for exploration of the SVG, and (3) the *javascript* library to support these files being mounted on the web.

Demonstration of how a blind user can explore the graph by “walking” a tree-like structure will be given. A key enhancement is the ability to explore the content at different levels of understanding; the user chooses to hear either the bare basic factual description or a more descriptive layer of feedback that can offer the user insight.

Bayesian social network analysis with Bergm

Alberto Caimo

Dublin Institute of Technology

Keywords: Bayesian analysis, Exponential random graph models, Monte Carlo methods

Webpages: https://CRAN.R-project.org/package=Bergm

Exponential random graph models (ERGMs) are a very important family of statistical models for analyzing network data. From a computational point of view, ERGMs are extremely difficult to handle since their normalising constant, which depends on model parameters, is intractable. In this talk, we show how parameter inference can be carried out in a Bayesian framework using MCMC strategies which circumvents the need to calculate the normalising constants.

The new version of the **Bergm** package for *R* (Caimo and Friel 2014) provides a comprehensive framework for Bayesian analysis for ERGMs useing the approximate exchange algorithm (Caimo and Friel 2011) and calibration of the pseudo-posterior distribution (Bouranis, Friel, and Maire 2015) to sample from the ERGM parameter posterior distribution. The package can also supply graphical Bayesian goodness-of-fit procedures that address the issue of model adequacy.

This talk will have a strong focus on the main practical implementation features of the software that will be described by the analysis of real network data (with various applications in Neuroscience and Organisation Science).

References

Bouranis, L., N. Friel, and F. Maire. 2015. “Bayesian Inference for Misspecified Exponential Random Graph Models.” *arXiv Preprint arXiv:1510.00934*.

Caimo, A., and N. Friel. 2011. “Bayesian Inference for Exponential Random Graph Models.” *Social Networks* 33 (1): 41–55.

———. 2014. “Bergm: Bayesian Exponential Random Graphs in R.” *Journal of Statistical Software* 61 (2): 1–25.

The **renjin** package: Painless Just-in-time Compilation for High Performance R

Alexander Bertram[1]

1. BeDataDriven

Keywords: performance, compliation, Renjin

Webpages: http://docs.renjin.org/en/latest/package/

R is a highly dynamic language that has developed, in some circles, a reputation for poor performance. New programmers are counseled to avoid `for` loops and experienced users condemened to rewrite perfectly good R code in C++.

Renjin is an alternative implementation of the R language that includes a Just-in-Time compiler which uses information at runtime to dynamically specialize *R* code and generate highly-efficient machine code, allowing users to write "normal", expresssive R code and let the compiler worry about performance.

While Renjin aims to provide a complete alternative to the GNU R interpreter, it is not yet fully compatible with all R packages, and lacks a number of features, including graphics support. For this reason, we present **renjin**, a new package that embeds Renjin's JIT compiler in the existing GNU R compiler, enabling even novice programmers to achieve a high performance without resorting to C++ or making the switch to a different interpreter.

This talk will introduce the techniques behind Renjin's optimizing compiler, demonstrate how it can be simply applied to performance-critical sections of R code, and some tips and tricks for getting the most of out of **renjin**.

R-Ladies Global Community

Alice Daish[1,2], Hannah Frick[1,3], Gabriela de Queiroz[1,4], Erin LeDell[1,5], Chiin-Rui Tan[1], Claudia Vitolo[1,6]

1. *R-Ladies Global Leadership*
2. *The British Museum*
3. *Mango Solutions*
4. *SelfScore*
5. *H2O.ai*
6. *European Centre for Medium-Range Weather Forecasts*

Keywords: R-Ladies, Diversity, R community

Webpages: https://rladies.org/

The *R* community suffers from an underrepresentation of women* in every role and area of participation: whether as leaders (no women on the *R* core team, 5 of 37 female ordinary members of the *R*-Foundation), package developers (around 10% women amongst CRAN maintainers, Forwards Task Force 2017; Mair et al. 2015), conference speakers and participants (around 28% at useR! 2016, Forwards Task Force 2017), educators, or users.

As a diversity initiative alongside the Forwards Task Force, *R*-Ladies' mission is to achieve proportionate representation by encouraging, inspiring, and empowering the minorities currently underrepresented in the *R* community. *R*-Ladies' primary focus is on supporting the *R* enthusiasts who identify as an underrepresented gender minority to achieve their programming potential, by building a collaborative global network of *R* leaders, mentors, learners, and developers to facilitate individual and collective progress worldwide.

Since *R*-Ladies Global was created a year ago we have grown exponentially to more than 4000 *R*-Ladies in 15 countries and have established a great brand. We want to share the amazing work *R*-Ladies has achieved, future plans and how the *R* community can support and champion *R*-Ladies around the world.

References

Forwards Task Force. 2017. https://forwards.github.io/data.

Mair, Patrick, Eva Hofmann, Kathrin Gruber, Reinhold Hatzinger, Achim Zeileis, and Kurt Hornik. 2015. "Motivation, Values, and Work Design as Drivers of Participation in the R Open Source Project for Statistical Computing." *Proceedings of the National Academy of Sciences* 112 (48): 14788–92. doi:10.1073/pnas.1506047112.

Neural Embeddings and NLP with R and Spark

Ali Zaidi
Microsoft, Algorithms and Data Science; Stanford University

Keywords: NLP, Spark, Deep Learning, Network Science

Webpages: https://github.com/akzaidi

Neural embeddings (Bengio et al. (2003), Olah (2014)) aim to map words, tokens, and general compositions of text to vector spaces, which makes them amenable for modeling, visualization, and inference. In this talk, we describe how to use neural embeddings of natural and programming languages using *R* and *Spark*. In particular, we'll see how the combination of a distributed computing paradigm in *Spark* with the interactive programming and visualization capabilities in *R* can make exploration and inference of natural language processing models easy and efficient.

Building upon the tidy data principles formalized and efficiently crafted in Wickham (2014), Silge and Robinson (2016) have provided the foundations for modeling and crafting natural language models with the `tidytext` package. In this talk, we'll describe how we can build scalable pipelines within this framework to prototype text mining and neural embedding models in *R*, and then deploy them on *Spark* clusters using the `sparklyr` and the `RevoScaleR` packages.

To describe the utility of this framework we'll provide an example where we'll train a sequence to sequence neural attention model for summarizing git commits, pull request and their associated messages (Zaidi (2017)), and then deploy them on *Spark* clusters where we will then be able to do efficient network analysis on the neural embeddings with a `sparklyr` extension to `GraphFrames`.

References

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. "A Neural Probabilistic Language Model." *J. Mach. Learn. Res.* 3 (March). JMLR.org: 1137–55. http://dl.acm.org/citation.cfm?id=944919.944966.

Olah, Christopher. 2014. "Deep Learning, NLP, and Representations." https://colah.github.io/posts/2014-07-NLP-RNNs-Representations/.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1 (3). The Open Journal. doi:10.21105/joss.00037.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1): 1–23. doi:10.18637/jss.v059.i10.

Zaidi, Ali. 2017. "Summarizing Git Commits and Github Pull Requests Using Sequence to Sequence Neural Attention Models." CS224N: Final Project, Stanford University.

R4ML: A Scalable R for Machine Learning

Alok Singh[1]

1. Spark Technology Center, IBM

Keywords: R, Distributed/Scalable, Machine Learning, SparkR, SystemML

Webpages: https://github.com/SparkTC/R4ML

R is the de facto standard for statistics and analysis. In this talk, we introduce R4ML, a new open-source *R* package for scalable machine learning from IBM. R4ML provides a bridge between *R*, Apache SystemML and SparkR, allowing *R* scripts to invoke custom algorithms developed in SystemML's R-like domain specific language. This capability also provides a bridge to the algorithm scripts that ship with Apache SystemML, effectively adding a new library of prebuilt scalable algorithms for *R* on Apache Spark. R4ML integrates seamlessly SparkR, so data scientists can use the best features of SparkR and SystemML together in the same script. In addition, the R4ML package provides a number of useful new scalable *R* functions that simplify common data cleaning and statistical analysis tasks.

Our talk will begin with an overview of the R4ML package, its API, supported canned algorithms, and the integration to Spark and SystemML. We will walk through a small example of creating a custom algorithm and a demo of canned algorithm. We will share our experiences using R4ML technology with IBM clients. The talk will conclude with pointers to how the audience can try out R4ML and discuss potential areas of community collaboration.

Curve Linear Regression with **clr**

Amandine Pierrot[1], Yannig Goude[1,2] and Qiwei Yao[3]

1. EDF R&D, France
2. Paris-Sud University, France
3. London School of Economics, UK

Keywords: Dimension reduction, Correlation dimension, Singular value decomposition, Load forecasting

We present a new *R* package for curve linear regression: the **clr** package.

This package implements a new methodology for linear regression with both curve response and curve regressors, which is described in Cho et al. (2013) and Cho et al. (2015).

The key idea behind this methodology is dimension reduction based on a singular value decomposition in a Hilbert Space, which reduces the curve regression problem to several scalar linear regression problems.

We apply curve linear regression with **clr** to model and forecast daily electricity loads.

References

Bathia, N., Q. Yao, and F. Ziegelmann. 2010. "Identifying the Finite Dimensionality of Curve Time Series." *The Annals of Statistics* 38: 3352–86.

Cho, H., Y. Goude, X. Brossat, and Q. Yao. 2013. "Modelling and Forecasting Daily Electricity Load Curves: A Hybrid Approach." *Journal of the American Statistical Association* 108: 7–21.

———. 2015. "Modelling and Forecasting Daily Electricity Load via Curve Linear Regression." In *Modeling and Stochastic Learning for Forecasting in High Dimension*, edited by Anestis Antoniadis and Xavier Brossat, 35–54. Springer.

Fan, J., and Q. Yao. 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer.

Hall, P., and J. L. Horowitz. 2007. "Methodology and Convergence Rates for Functional Linear Regression." *The Annals of Statistics* 35: 70–91.

useR! 2017 Oral Presentation

Clustering transformed compositional data using *coseq*

Andrea Rau[1], *Antoine Godichon-Baggioni*[2], *and Cathy Maugis-Rabusseau*[2]

[1] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy en Josas, France

[2] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, 31400 Toulouse, France

Abstract: Although there is no shortage of clustering algorithms proposed in the literature, the question of the most relevant strategy for clustering compositional data (i.e., data made up of profiles, whose rows belong to the simplex), remains largely unexplored, particularly in cases where the observed value of an observation is equal or close to zero for one or more samples. This work is motivated by the analysis of two sets of compositional data, both focused on the categorization of profiles but arising from considerably different applications: (1) identifying groups of co-expressed genes from high-throughput RNA sequencing data, in which a given gene may be completely silent in one or more experimental conditions; and (2) finding patterns in the usage of stations over the course of one week in the Velib' bike sharing system in Paris, France. For both of these applications, we propose the use of appropriate data transformations in conjunction with either Gaussian mixture models or K-means algorithms and penalized model selection criteria. Using our Bioconductor package *coseq*, we illustrate the user-friendly implementation and visualization provided by our proposed approach, with a focus on the functional coherence of the gene co-expression clusters and the geographical coherence of the bike station groupings.

Keywords: Clustering, compositional data, K-means, mixture model, transformation, co-expression

The use of R in predictive maintenance: A use case with TRUMPF Laser GmbH

Andreas Prawitt

eoda GmbH

Keywords: data science, predictive maintenance, industry 4.0, business, industry, use case,

The buzz for industry 4.0 continues – digitalizing business processes is one of the main aims of companies in the 21st century. One topic gains particular importance: predictive maintenance. Enterprises use this method in order to cut production and maintenance costs and to increase reliability.

Being able to predict machine failures, performance drops or quality deterioration is a huge benefit for companies. With this knowledge, maintenance and failure costs can be reduced and optimized.

With the help of R and its massive community, analysts can apply the best algorithms and methods for predictive maintenance. When a good analytic model for predictive maintenance has been found, companies are challenged to implement them in their own environments and workflows. Especially regarding the workflow across different departments, it is necessary to find an appropriate solution which is capable of interdisciplinary work, as well.

My talk will show how this challenge was solved for TRUMPF Laser GmbH, a subsidiary of TRUMPF, a world-leading high-technology company which offers production solutions in the machine tool, laser and electronic sectors. I would like to share my experience with R and predictive maintenance in a real-world industry scenario and show the audience how to automate R code and visualize it in a front-end solution for all departments involved.

useR! 2017 Oral Presentation

Updates to the Documentation System for R

Andrew Redd[1]

1. 1. Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City , UT

Funding: This work is supported by funding from the R Consortium and The University of Utah Center for Clinical and Translational Science (NIH 5UL1TR001067-02).

Abstract: Over the last few years while the open source statistical package R has come to prominence it has gained important resources, such as multiple flexible class systems. However, methods for documentation have not kept pace with other advances in the language. I will present the work of the R Documentation Task Force, an R Consortium Working Group, in creating the next generation of documentation system for R.

The new documentation system is based off a S4 formal class system and exists independent of but is complimentary to the packaging system in R. Documentation objects are stored as objects and as such can be manipulated programmatically as with all R objects.

This approach creates a "many in-many out" approach, meaning that developers of software and documentation can create documentation in the format that is easiest for them, such as Rd or Roxygen, and users of the documentation can read or utilize documentation in a convenient format. Since R also makes use of code from other languages such as C++, this creates faculties for including documentation without recreating it.

This work is based on input from the R Documentation Task Force, which is a working group, supported by the R Consortium and the University of Utah Center for Clinical and Translational Science, consisting of R Core developers, representatives from the R Consortium member companies and community developers with relevant interest in documentation.

Good documentation is critical for researchers to disseminate computational research methods, either internally or externally to their organization. This work will facilitate the creation of documentation by making documentation immediately accessible and promote documentation consumption through multiple outputs which can be implemented by developers.

useR! 2017 Oral Presentation

Can you keep a secret?

Andrie de Vries[1] and Gábor Csárdi[2]

1. Senior Programme Manager, Algorithms and Data Science, Microsoft
2. Independent consultant

Keywords: Asymmetric encryption, Public key encryption

When you use *R* to connect to a database, cloud computing service or other API, you must supply passwords, for example database credentials, authentication keys, etc.

It is easy to inadvertently leak your passwords and other secrets, e.g. accidentally adding your secrets to version control or logs.

A new package, **secret** solves this problem by allowing you to encrypt and encrypt secrets using public key encryption. The package is available at github [@secret] and soon also on CRAN.

If you attend this session, you will learn:

- Patterns for inadvertently leak secrets
- The essentials of public key cryptography: how to create an asymmetric key pair (public and private key)
- How to create a vault with encrypted secrets using the **secret** package
- How to share these secrets with your collaborators by encrypting the secret with their public key
- How you can do all of this in 5 lines of R code

This session will appeal to all R users who must use passwords to connect to services.

When is an Outlier an Outlier? The O3 plot.

Antony Unwin

Whether a case might be identified as an outlier depends on the other cases in the dataset and on the variables available. A case can stand out as unusual on one or two variables, while appearing middling on the others. If a case is identified as an outlier, it is useful to find out why. This paper introduces a new display, the O3 plot (Overview Of Outliers), for supporting outlier analyses, and describes its implementation in R.

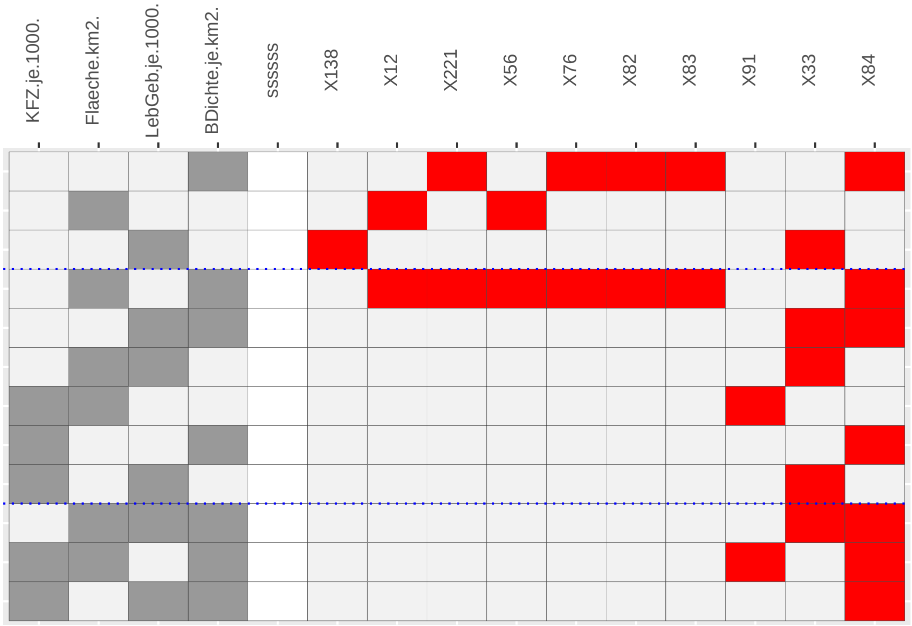


Figure 1 shows an example of an O3 plot for four German demographic variables recorded for the 299 Bundestag constituencies. There is a row for each variable combination for which outliers were found and two blocks of columns. Each row of the block on the left shows which variable combination defines that row. There are 4 variables, so there are 4 columns, one for each variable, and a cell is coloured grey if that variable is part of the combination. The combinations (the rows) are sorted by numbers of outliers found within numbers of variables in the combination, and blue dotted lines separate the combinations with different numbers of variables. The columns in the left block are sorted by how often the variables occur. A boundary column separates this block from the block on the right that records the outliers found with whichever outlier identification algorithm was used (in this case Wilkinson's HDoutliers with alpha=0.05). There is one column for each case that is found to be an outlier at least once and these columns are sorted by the numbers of times the cases are outliers.

Given n cases and p variables there would be $(p + 1 + n)$ columns if all cases were an outlier on some combination of variables. And if outliers were identified for all possible combinations there would be $2^p - 1$ rows. An O3 plot has too many rows if there are lots of variables with many combinations having outliers and it has too many columns if there are lots of cases identified as outliers on at least one variable combination. Combinations are only reported if outliers are found for them and cases are only reported which occur at least once as an outlier.

O3 plots show which cases are identified often as outliers, which are identified in single dimensions, and which are only identified in higher dimensions. They highlight which variables and combinations of variables may be affected by possible outliers.

jug: Building Web APIs for R

Bart Smeets[1]

1. dataroots

Keywords: REST, API, web, http

Webpages: https://CRAN.R-project.org/package=jug, https://github.com/Bart6114/jug

jug is a web framework for *R*. The framework helps to easily set up API endpoints. Its main goal is to make building and configuration of web APIs as easy as possible, while still allowing in-depth control over HTTP request processing when needed.

A **jug** instance allows one to expose solutions developed in R to the web or and/or applications communicating over HTTP. This way, other applications can gain access to, e.g. custom *R* plotting functions or generate new predictions based on a trained machine learning model.

jug is build upon **httpuv**. This results in a stable and robust back-end. Recently, endeavors have been made to allow a **jug** instance to process requests in parallel. The GitHub repository includes a Dockerfile to ease productionisation and containerisation of a **jug** instance.

During this talk, a tangible reproducible example of creating an API based on a machine learning model will be presented and some of the challenges and experiences in exposing R based results through an API will be discussed.

How to Use (R)Stan to Estimate Models in External R Packages

Ben Goodrich [1]

1. *Columbia University*

Keywords: Bayesian, developeRs

Webpages: rstan, rstanarm, rstantools, StanHeaders, bayesplot, shinyStan, loo, home page

The **rstan** package provides an interface from *R* to the Stan libraries, which makes it possible to access Stan's advanced algorithms to draw from any posterior distribution whose density function is differentiable with respect to the unknown parameters. The **rstan** package is ranked in the 99-th percentile overall on Depsy due to its number of downloads, citations, and use in other projects. This talk is a follow-up to the very successful Stan workshop at useR2016 and will be more focused on how maintainers of other *R* packages can easily use Stan's algorithms to estimate the statistical models that their packages provide. These mechanisms were developed to support the **rstanarm** package for estimating regression models with Stan and have since been used by over twenty *R* packages, but they are perhaps not widely known and difficult to accomplish manually. Fortunately, the `rstan_package.skeleton` function in the **rstantools** package can be used to automate most of the process, so the package maintainer only needs to write the log-posterior density (up to a constant) in the Stan language and provide an *R* wrapper to call the pre-compiled *C++* representation of the model. Methods for the resulting *R* object can be defined that allow the user to analyze the results using post-estimation packages such as **bayesplot**, **ShinyStan**, and **loo**.

References

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. doi:10.18637/jss.v076.i01.

useR! 2017 Oral Presentation

Text Analysis and Text Mining Using R

Kenneth Benoit

Keywords: text analysis, text mining, machine learning, social media

Summary

A useR! Talk about text analysis and text mining using R. I would cover the broad set of tools for text analysis and natural language processing in R, with an emphasis on my R package **quanteda** but also covering other major tools in the R ecosystem for text analysis (e.g. **stringi**).

The talk would is tutorial covers how to perform common text analysis and natural language processing tasks using R. Contrary to a belief popular among some data scientists, when used properly, R is a fast and powerful tool for managing even very large text analysis tasks. My talk would present the many option available, demonstrate that these work on large data, and compare the features of R for these tasks versus popular options in Python.

Specifically, I will demonstrate how to format and input source texts, how to structure their metadata, and how to prepare them for analysis. This includes common tasks such as tokenisation, including constructing ngrams and "skip-grams", removing stopwords, stemming words, and other forms of feature selection. I will also show to how to tag parts of speech and parse structural dependencies in texts. For statistical analysis, I will show how R can be used to get summary statistics from text, search for and analyse keywords and phrases, analyse text for lexical diversity and readability, detect collocations, apply dictionaries, and measure term and document associations using distance measures. Our analysis covers basic text-related data processing in the R base language, but most relies on the **quanteda** package (https://github.com/kbenoit/quanteda) for the quantitative analysis of textual data. We also cover how to pass the structured objects from quanteda into other text analytic packages for doing topic modelling, latent semantic analysis, regression models, and other forms of machine learning.

About me

Kenneth Benoit is Professor of Quantitative Social Research Methods at the London School of Economics and Political Science. His current research focuses on automated, quantitative methods for processing large amounts of textual data, mainly political texts and social media. Current interest span from the analysis of big data, including social media, and methods of text mining. For the past 5 years, he has been developing a major R package for text analysis, **quanteda**, as part of European Research Council grant ERC-2011-StG 283794-QUANTESS.

Analysis of German Fuel Prices with R

Boris Vaillant

Quantitative Consulting, Bonn

Keywords: Analytics, Marketing, **tidyverse**, **purrr**, **ggplot2**, **rgdal**, **sp** and more
Webpages: https://creativecommons.tankerkoenig.de (sic), https://www.openstreetmap.org/

We present an R-based analysis to measure the impact of different market drivers on fuel prices in Germany. The analysis is based on the open dataset on German fuel prices, bringing in many additional open data sets along the way.

- **Overview of the dataset**
  1. History, Legal framework and data collection

  2. Current uses in "price-finder apps"
  3. Structure of the dataset

  4. Preparation of the data
  5. A first graphical analysis
  – price levels
  – weekly and daily pricing patterns
- **Overview of potential price drivers and corresponding data sources**
  1. A **Purrr** workflow for preparing regional data from Destatis
  – Number of registered cars
  – Number of fuel stations
  – Number of inhabitants
  – Mean income, etc.
  2. Determining geographical market drivers with OSM data using **sp**, **rgdal**, **geosphere**
  – Branded vs independent
  – Location: higwhway, close to highway exit ("Autohof") etc.
  – Proximity to competitors, etc.
  3. Cost drivers
  – Market prices for crude oil
  – Distance of fuel station to fuel depot
  – Land lease and property-prices
  4. Outlook:
  – Weather
  – Traffic density

Based on this data, we will present different modelling approaches to quantify the impact of the above drivers on average price levels. We will also give an outlook and first results on temporal pricing patterns and indicators for competitive or anti-competitive behaviour.

This talk is a condensed version of an online R-workshop that I am currently preparing and which I expect to be fully available at the time of UseR 2017.

Dynamic Assessment of Microbial Ecology (*DAME*): A Shiny App for Analysis and Visualization of Microbial Sequencing Data

Brian D Piccolo[1,2], Umesh D Wankhade[1,2], Sree V Chintapalli[1,2], and Kartik Shankar[1,2]

1. Arkansas Children's Nutrition Center, Little Rock, AR, USA
2. Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Keywords: Shiny, microbiome, sequencing, ecology, 16S rRNA

Webpages: https://acnc-shinyapps.shinyapps.io/DAME/, https://github.com/bdpiccolo/ACNC-DAME

A new renaissance in knowledge about the role of commensal microbiota in health and disease is well underway facilitated by culture-independent sequencing technologies; however, microbial sequencing data poses new challenges (e.g., taxonomic hierarchy, overdispersion) not generally seen in more traditional sequencing outputs. Additionally, complex study paradigms from clinical or basic research studies necessitate a multilayered analysis pipeline that can seamlessly integrate both primary bioinformatics and secondary statistical analysis combined with data visualization.

In order to address this need, we created a web-based Shiny app, titled DAME, which allows users not familiar with *R* programming to import, filter, and analyze microbial sequencing data from experimental studies. DAME only requires two files (a BIOM file with sequencing reads combined with taxonomy details, and a csv file containing experimental metadata), which upon upload will trigger the app to render a linear work-flow controlled by the user. Currently, DAME supports group comparisons of several ecological estimates of α-diversity (ANOVA) and β-diversity indices (ordinations and PERMANOVA). Additionally, pairwise differential comparisons of operational taxonomic units (OTUs) using Negative Binomial Regression at all taxonomic levels can be performed. All analyses are accompanied by dynamic graphics and tables for complete user interactivity. DAME leverages functions derived from **phyloseq**, **vegan**, and **DESeq2** packages for microbial data organization and analysis and **DT**, **highcharter*** and **scatterD3** for table and plot visualizations. Downloadable options for α-diversity measurements and **DESeq2** table outputs are also provided.

The current release (v0.1) is available online (https://acnc-shinyapps.shinyapps.io/DAME/) and in the Github repository (https://github.com/bdpiccolo/ACNC-DAME). *This app uses Highsoft software with non-commercial packages. Highsoft software product is not free for commercial use. Funding supported by United States Department of Agriculture-Agricultural Research Service Project: 6026-51000-010-05S.

Hosting Data Packages via `drat`: A Case Study with Hurricane Exposure Data

Brooke Anderson[1,*], **Dirk Eddelbuettel**[2,3]

1. Department of Environmental & Radiological Health Sciences, Colorado State University, Fort Collins, CO, USA
2. Debian and R Projects
3. Ketchum Trading, Chicago, IL, USA

*Contact author: brooke.anderson@colostate.edu

Keywords: drat repositories; package repositories; data packages; reproducible research

R packages offer the chance to distribute large datasets while also providing functions for exploring and working with that data. However, data packages often exceed the suggested size of CRAN packages, which is a challenge for package maintainers who would like to share their code through this central and popular repository. In this talk, we outline an approach in which the maintainer creates a smaller code package with the code to interact with the data, which can be submitted to CRAN, and a separate data package, which can be hosted by the package maintainer through a personal **drat** repository. Although **drat** repositories are not mainstream, and so cannot be listed with an "Includes" or "Depends" dependency for a package submitted to CRAN, we suggest a way of including the data package as a suggested package and incorporating conditional code in the executable code within vignettes, examples, and tests, as well as conditioning functions in the code package to check for the availability of the data package. We illustrate this approach for a pair of packages , **hurricaneexposure** and **hurricaneexposuredata**, that allows users to explore exposure to hurricanes and tropical storms in the United States. This approach may prove useful for a number of R package maintainers, especially with the growing trend to the sharing and use of open data in many of the fields in which R is popular.

Collaborative Development in R: A Case Study with the **sparsebn** package

Bryon Aragam[1]

1. University of California, Los Angeles

Webpages: https://CRAN.R-project.org/package=sparsebn

With the massive popularity of R in the statistics and data science communities along with the recent movement towards open development and reproducible research with CRAN and GitHub, R has become the de facto go-to for cutting edge statistical software. With this movement, a problem faced by many groups is how individual programmers can work on related codebases in an open, collaborative manner while emphasizing good software practices and reproducible research. The **sparsebn** package, recently released on CRAN, is an example of this dilemma: **sparsebn** is a family of packages for learning graphical models, with different algorithms tailored for different types of data. Although each algorithm shares many similarities, different researchers and programmers were in charge of implementing different algorithms. Instead of releasing disparate, unrelated packages, our group developed a shared family of packages in order to streamline the addition of new algorithms so as to minimize programming overhead (the dreaded "data munging" and "plumbing" work). In this talk, I will use **sparsebn** as a case study in collaborative research and development, illustrating both the development process and the fruits of our labour: A fast, modern package for learning graphical models that leverages cutting-edge trends in high-dimensional statistics and machine learning.

rags2ridges: A One-Stop-Go for Network Modeling of Precision Matrices

Carel F.W. Peeters[1], Anders E. Bilgrau[2], and Wessel N. van Wieringen[1,3]

1. Dept. of Epidemiology & Biostatistics, VU University medical center, Amsterdam, The Netherlands
2. Biostatistics Unit, Novo Nordisk, Aalborg, Denmark
3. Dept. of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

Keywords: Data integration, Graphical modeling, High-dimensional precision matrix estimation; Networks

Webpages: https://CRAN.R-project.org/package=rags2ridges, https://github.com/CFWP/rags2ridges

Contact: cf.peeters@vumc.nl

A contemporary use for inverse covariance matrices (aka precision matrices) is found in the data-based reconstruction of networks through graphical modeling. Graphical models merge probability distributions of random vectors with graphs that express the conditional (in)dependencies between the constituent random variables. The **rags2ridges** package enables L2-penalized (i.e., ridge) estimation of the precision matrix in settings where the number of variables is large relative to the sample size. Hence, it is a package where high-dimensional (HD) data meets networks.

The talk will give an overview of the **rags2ridges** package. Specifically, it will show that the package is a one-stop-go as it provides functionality for the extraction, visualization, and analysis of networks from HD data. Moreover, it will show that the package provides a basis for the vertical (across data sets) and horizontal (across platforms) integration of HD data stemming from omics experiments. Last but not least, it will explain why many rap musicians are stating that one should 'get *ridge*, or die trying'.

References

https://arxiv.org/abs/1509.07982

https://arxiv.org/abs/1608.04123

http://dx.doi.org/10.1016/j.csda.2016.05.012

useR! 2017 Oral Presentation

countreg: Tools for count data regression

Christian Kleiber[1] *and Achim Zeileis*[2]

1. *Universität Basel, Switzerland*
2. *Universität Innsbruck, Austria*

Keywords: Count data regression, model diagnostics, rootogram, visualization

Webpages: https://R-Forge.R-project.org/projects/countreg

The interest in regression models for count data has grown rather rapidly over the last 20 years, partly driven by methodological questions and partly by the availability of new data sets with complex features (see, e.g., Cameron and Trivedi 2013). The **countreg** package for *R* provides a number of fitting functions and new tools for model diagnostics. More specifically, it incorporates enhanced versions of fitting functions for hurdle and zero-inflation models that have been available via the **pscl** package for some 10 years (Zeileis, Kleiber, and Jackman 2008), now also permitting binomial responses. In addition, it provides zero-truncation models for data without zeros, along with **mboost** family generators that enable boosting of zero-truncated and untruncated count data regressions, thereby supplementing and extending family generators available with the **mboost** package. For visualizing model fits, **countreg** offers rootograms (Tukey 1972; Kleiber and Zeileis 2016) and probability integral transform (PIT) histograms. A (generic) function for computing (randomized) quantile residuals is also available. Furthermore, there are enhanced options for `predict()` methods. Several new data sets from a variety of fields (including dentistry, ethology, and finance) are included.

Development versions of **countreg** have been available from R-Forge for some time, a CRAN release is planned for summer 2017.

References

Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data.* 2nd ed. Cambridge: Cambridge University Press.

Kleiber, Christian, and Achim Zeileis. 2016. "Visualizing Count Data Regressions Using Rootograms." *The American Statistician* 70 (3): 296–303.

Tukey, John W. 1972. "Some Graphic and Semigraphic Displays." In *Statistical Papers in Honor of George W. Snedecor*, edited by T. A. Bancroft, 293–316. Ames, IA: Iowa State University Press.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. "Regression Models for Count Data in R." *Journal of Statistical Software* 27 (8): 1–25. http://www.jstatsoft.org/v27/i08/.

Diversity of the R Community

Julie Josse[1] and the R taskforce on women

1. Ecole Polytechnique, France

Keywords: R foundation, R community, gender gap, diversity, useR! conferences

Webpages: https://forwards.github.io/

R Forwards is a R Foundation taskforce which aims at leading the R community forwards in widening the participation of women and other under-represented groups. We are organized in sub-teams that work on specific tasks, such as data collection and analysis, social media, gathering teaching materials, organizing targeted workshops, keep track of scholarships and interesting diversity initiatives, etc. In this talk, I will present an overview of our activities and in particular the work of the survey team who analyzed the questionnaire run at useR! 2016. We collected information on the participants socio-demographic, experiences and interest in R to get a better understanding of how to make the R community a more inclusive environment. We regularly post our results with blogs. Based on this analysis, I will present some of our recommendations.

Title Sports Betting and R: How R is changing the sports betting world

Marco Blume

Keywords: Sports Betting, Sports Analytics, Vegas, Markets

Webpages - https://cran.r-project.org/web/packages/odds.converter/index.html - https://cran.r-project.org/web/packages/pinnacle.API/index.html - http://pinnacle.com/

Sports Betting markets are one of the purest prediction markets that exist and are yet vastly misunderstood by the public. Many assume that the center of the sports betting world is situated in Las Vegas. However, in the modern era, sports bookmaking is a task that looks a lot like market making in finance with sophisticated algorithmic trading systems running and constantly adjusting prices in real-time as events occur. But, unlike financial markets, sports are governed by a set of physical rules and can usually be measured and understood. Since the late 90s, Pinnacle has been one of the largest sportsbooks in the world and one of the only sportsbooks who will take wagers from professional bettors (who win in the long term). Similar to card counters in Blackjack, most other sportsbook will ban these winners. At Pinnacle the focus is on modeling, automation, data science and R is a central piece of the business and a large number of customers use an API to interact with us. In this talk, we dispel common misconceptions about the sports betting world and show how this is actually a very sexy problem in modeling and data science and show how we are using R to try to beat Vegas and other sportsbooks every day in a form of data science warfare. Since the rise of in-play betting markets, an operator must make a prediction in real time on the probability of outcomes for the remainder of an event within a very small margin of error. Customers can compete by building their own models or utilizing information that might not be accounted for in the market and expressing their belief through wagering. Naturally, a customer will generally wager when they believe they have an edge, and then the operator must determine how to change its belief after each piece of new information (wagers, in-game events, etc). This essentially involves predicting how much information is encoded in a wager, which depends partially on the sharpness of each customer, and then determining how to act on that information to maximize profits. One way to look at this is that we are aggregating, in a smart way, the world's models, opinions, and information when we come up with a price. This is a powerful concept and is why, for example, political prediction markets are much more accurate than polls or pundits. For this reason, we are releasing another package to CRAN very soon: We will be releasing a package that has all our odds for the entire MLB season 2016 and can be combined with the very popular Lahman package to build predictive models and to measure the prediction vs real market data to see how your model would have performed in a real market. We believe this is a very exciting (and difficult) problem to use for educational purposes. This package can be used in conjunction with two of our existing packages already on CRAN for a few years: odds.converter (to convert between betting market odds types and probabilities) and Pinnacle.API (used to interact with Pinnacle's real-time odds API in R).

Even if you have no interest in sports or wagering, we believe this is a fascinating problem and our data and tools are perfect for the R community at large to work with, for academic reasons or for hobby.

Link2GI - Easy linking Open Source GIS with R

Christoph Reudenbach [1], Florian Detsch [2] and Tim Appelhans [3]

1. *University of Marburg | GIS and Environmental Modeling*
2. *University of Marburg | Environmental Informatics*
3. *GfK Geomarketing | Nürnberg | Germany*

Keywords Spatial analysis, Setup, GRASS, SAGA, OTB, QGIS

Despite the well known capabilities of spatial analysis and data handling in the world of *R*, an enormous gap persists between *R* and the mature open source Geographic Information System (GIS) and Remote Sensing (RS) software community. Prominent representatives like *QGIS*, *GRASS GIS* and *SAGA GIS* provide comprehensive and continually growing collections of highly sophisticated algorithms that are mostly fast, stable and usually well proofed by the community

Although a number of *R* wrappers aim to bridge this gap (eg **rgrass7** for *GRASS GIS 7.x*, **RSAGA** for *SAGA GIS*) – among which **RQGIS** is the most recent outcome to realize a simple access to the powerful *QGIS* command line interface – most of these packages are not that easy to setup. Most of the wrappers are trying to find and/or set an appropriate environment, nevertheless it is in many cases at least cumbersome to get all necessary settings correct, especially if one has to work with restricted rights or parallel installations of the same GIS software.

In order to overcome known limitations, the package **link2GI** provides a small framework for easy linking of *R* to major GIS software. Here, linking simply means to provide all necessary environment settings as well as full access to the command line APIs of these software tools, whereby the strategy differs from software to software. As a result an easy entrance door for linking current versions of *GRASS7.x GIS*, *SAGA GIS*, *QGIS* as well as other command line tools like the *Orfeo Toolbox* (OTB) to *R* is provided. The package focus on both *R* users that are not very familiar with the conditions and pitfalls of their preferred operating system and more experienced users that want to have some comfortable shortcuts for a seamless integration of e.g. *GRASS*. The most simple call `link2GI::linkGRASS7(x=anySpatialObject)` will search for the OS dependent installations of *GRASS 7*. Furthermore, it will setup the rsession according to the provided spatial object. All steps can be influenced manually which will significantly speed up the process. Especially if you work with already established *GRASS* databases it provides a convenient way to link mapsets and locations correctly.

The package is also providing some basic tools beyond simple linking. Since Edzer Pebesma's new **sf** package, it is for the first time possible to deal with big vector data sets (> 1.000.000 polygons or 25.000.000 vertices). Nevertheless it is advantagous to process the more sophisticeded spatial analysis with external GIS software. To improve this process **link2GI** provides a first version of direct reading and writing GRASS and SAGA vector data from and to R to speed up the conversion process. Finally, a first version of a common Orfeo Toolbox wrapper for simplifying *OTB* calls is introduced.

A restricted composite likelihood approach to modelling Gaussian geostatistical data

Mutambanengwe CK[1,2] and Faes C[2] and Aerts M[2]

1. Open Analytics NV, Antwerp, Belgium
2. IBiostat, Hasselt University, Hasselt, Belgium

Keywords: composite likelihood, effective sample size, REML, spatial dependence

Composite likelihood methods have become popular in spatial statistics. This is mainly due to the fact that large matrices need to be inverted in full maximum likelihood and this becomes computationally expensive when you have a large number of regions under consideration. We introduce restricted pairwise composite likelihood (RECL) methods for estimation of mean and covariance parameters in a spatial Gaussian random field, without resorting back to the full likelihood. A simulation study was carried out to investigate how this method works in settings of increasing domain as well as infill asymptotics, whilst varying the strength of correlation, with similar scenarios as Curriero and Lele (1999). Preliminary results showed that pairwise composite likelihoods tend to underestimate the variance parameters, especially when there is high correlation, while RECL corrects for the underestimation. Therefore, RECL is recommended if interest is in both the mean and the variance parameters. The methods are made available in the **spatialRECL** package and implemented in *R*. The methodology will be highlighted in the first part of the presentation, and some analysis will be made on a real data example of TSH levels from Galicia, Spain.

References

Curriero, F, and S Lele. 1999. "A Composite Likelihood Approach to Semivariogram Estimation." *J Agric Biol Envir S* 4 (1): 9–28.

A Benchmark of Open Source Tools for Machine Learning from R

Szilard Pafka
Epoch

Keywords: machine learning, predictive modeling, predictive accuracy, scalability, speed

Webpages: https://github.com/szilard/benchm-ml

Binary classification is one of the most widely used machine learning methods in business applications. If the number of features is not very large (sparse), algorithms such as random forests, gradient boosted trees or deep learning neural networks (and ensembles of those) are expected to perform the best in terms of accuracy. There are countless off-the-shelf open source implementations for the previous algorithms (e.g. R packages, Python scikit-learn, H2O, xgboost, Spark MLlib etc.), but which one to use in practice? Surprisingly, there is a huge variation between even the most commonly used implementations of the same algorithm in terms of scalability, speed, accuracy. In this talk we will see which open source tools work reasonably well on larger datasets commonly encountered in practice. Not surprizingly, all the best tools are available seamlessly from R.

Dynamic modeling and parameter estimation with dMod

Daniel Kaschek[1], Wolfgang Mader[1], Mirjam Fehling-Kaschek[1], Marcus Rosenblatt[1] and Jens Timmer[1]

1. University of Freiburg, Germany

Keywords: Parameter Estimation, ODEs, Systems Biology, Maximum-Likelihood, Profile-Likelihood

Webpages: https://github.com/dkaschek/dMod, https://github.com/dkaschek/cOde

ODE models to describe and understand interactions in complex dynamical systems are widely used in the physical sciences and beyond. In many situations, the model equations depend on parameters. When parameters are not known from first principle, they need to be estimated from experimental data.

The **dMod** package for *R* provides a framework for formulating complex reaction networks and estimating the inherent reaction parameters from experimental data. By design, different experimental conditions as well as explicit or implicit equality constraints, e.g., steady-state constraints, are formulated by parameter transformations which thereby take a central role in **dMod**. Since, in general, the observed reaction dynamics is a non-linear function of the reaction parameters, profile-likelihood methods are implemented to assess non-linear parameter dependencies and estimate parameter- and prediction confidence intervals.

Here, we present the abilities and particularities of our modeling framework. The methods are illustrated based on a minimal systems biology example.

Scraping data with rvest and purrr

Max Humber[1]

1. Borrowell

Keywords: rvest, purrr, webscraping, fantasy, sports

Webpages: http://www.maxhumber.com

Really interesting data never actually lives inside of a tidy csv. Unless, of course, you think *Iris* or *mtcars* is super interesting. Interesting data lives outside of comma separators. It's unstructured, and messy, and all over the place. It lives around us and on poorly formatted websites, just waiting and begging to be played with.

Finding and fetching and cleaning your own data is a bit like cooking a meal from scratch—instead of microwaving a frozen TV dinner. Microwaving food is simple. It's literally one step: put thing in microwave. There is, however, no singular step to making a proper meal from scratch. Every meal is different. The recipe for making coconut curry isn't the same as the recipe for Brussels sprout tacos. But both require a knife and a frying pan!

In "Scraping data with **rvest** and **purrr**" I will talk through how to pair and combine **rvest** (the knife) and **purrr** (the frying pan) to scrape interesting data from a bunch of websites. This talk is inspired by a recent blog post that I authored for and was well received by the r-bloggers.com community.

---

rvest is a popular *R* package that makes it easy to scrape data from html web pages.

purrr is a relatively new package that makes it easy to write code for a single element of a list that can be quickly generalized to the rest of that same list.

useR! 2017 Oral Presentation

Markov-Switching GARCH Models in R: The MSGARCH Package

David Ardia[1,2] *and Keven Bluteau*[1] *and Kris Boudt*[3,4] *and Leopoldo Catania*[5] *and Brian Peterson*[6] *and Denis-Alexandre Trottier*[2]

1. *University of Neuchatel*
2. *Laval University*
3. *Vrije Universiteit Brussels*
4. *Vrije Universiteit Amsterdam*
5. *University of Rome "Tor Vergata"*
6. *DV Trading*

Keywords: GARCH, MSGARCH, Markov–switching, conditional volatility, risk management

Webpages: https://CRAN.R-project.org/package=MSGARCH, https://github.com/keblu/MSGARCH

Markov–switching GARCH models have become popular to model the structural break in the conditional variance dynamics of financial time series. In this paper, we describe the *R* package **MSGARCH** which implements Markov–switching GARCH–type models very efficiently by using *C++* object–oriented programming techniques. It allows the user to perform simulations as well as Maximum Likelihood and Bayesian estimation of a very large class of Markov–switching GARCH–type models. Risk management tools such as Value–at–Risk and Expected–Shortfall calculations are available. An empirical illustration of the usefulness of the *R* package **MSGARCH** is presented.

useR! 2017 Oral Presentation

Package ggiraph: a ggplot2 Extension for Interactive Graphics

David Gohel[1]

1. ArData

Keywords: visualization, interactive

Webpages: https://CRAN.R-project.org/package=ggiraph, https://davidgohel.github.io/ggiraph/

With rise of data visualisation, **ggplot2** and **D3.js** tools have become very popular these last years. The first is providing an high level library for data visualisation whereas the latter is providing a low level library for binding graphical elements in a web context.

The **ggiraph** package combines both tools. From a user point of view, it enables the production of interactive graphics from **ggplot2** objects by using their extension mechanism. It provides useful interactive capabilities such as tooltips and zoom/pan. Last but not least, graphical elements can be selected when a ggiraph object is embedded in a Shiny app: selection will be available as a reactive value. The interface is simple, flexible and does not requires effort to be integrated in R Markdown documents or Shiny applications.

In this talk I will introduce **ggiraph** and show examples of using it as a data visualisation tools in RStudio, Shiny applications and R Markdown documents.

Change Point Detection in Persistence Diagrams

David Letscher[1] *and Darrin Speegle*[1]

1. *Saint Louis University*

Keywords: TDA, Persistence, Wavelets, Change-Point Detection

Webpages: https://github.com/speegled/cpbaywave

Topological data analysis (TDA) offers a multi-scale method to represent, visualize and interpret complex data by extracting topological features using persistent homology. We will focus on persistence diagrams, which are a way of representing the persistent homology of a point cloud. At their most basic level, persistence diagrams can give something similar to clustering information, but they also can give information about loops or other topological structures within a data set.

Wavelets are another multi-scale tool used to represent, visualize and interpret complex data. Wavelets offer a way of examining the local changes of a data set while also estimating the global trends.

We will present two algorithms that combine wavelets and persistence. First, we use a wavelet based density estimator to bootstrap confidence intervals in persistence diagrams. Wavelets seem well-suited for this, since if the underlying data lies on a manifold, then the density should have discontinuities that will need to be detected. Additionally, the wavelet based algorithm is fast enough to allow some cross-validation of the tuning parameters. Second, we present an algorithm for detecting the most likely change point of the persistent homology of a time series.

The majority of this talk will consist of presenting examples which will illustrate persistence diagrams, the change point detection algorith, and the types of changes in geometric and/or topological structure in data that can be detected via this algorithm.

useR! 2017 Oral Presentation

We R What We Ask: The Landscape of R Users on Stack Overflow

David Robinson[1]

1. Stack Overflow

Keywords: r, data science, web traffic, visualization

Since its founding in 2008, the question and answer website Stack Overflow has been a valuable resource for the *R* community, collecting more than 175,000 questions about the *R* that are visited millions of times each month. This makes it a useful source of data for observing trends about how people use and learn the language. In this talk, I show what we can learn from Stack Overflow data about the global use of the *R* language over the last decade. I'll examine what ecosystems of *R* packages are used in combination, what other technologies are used alongside *R**, and what countries and cities have the highest density of users. Together, the data paints a picture of a global and rapidly growing community. Aside from presenting these results, I'll introduce interactive tools and visualizations that the company has published to explore this data, as well as a number of open datasets that analysts can use to examine trends in software development.

ompr: an alternative way to model mixed-integer linear programs

Dirk Schumacher

Keywords: integer programming, linear programming, modelling, optimization

Webpages: https://github.com/dirkschumacher/ompr

Many real world optimization problems, such as the popular traveling salesman problem, can be formulated as a mixed-integer linear program (MILP). The aim of MILP is to optimize a linear objective function, subject to a set of linear constraints. Over the past decades, specialized open-source and commercial solvers have been developed, such as the GNU Linear Programming Kit (GLPK) which can efficiently solve these kinds of problems.

In *R*, interfaces to these solvers are mostly matrix oriented. When solving a MILP in *R*, you would thus first need to develop your actual model and then translate it into code that constructs a matrix and vectors before passing it to a solver. Especially for more complex models, the *R* code might be rather hard to develop and to reason about without additional documentation.

ompr is a domain specific language that lets you model MILPs declaratively using functions like `set_objective`, `add_variable` or `add_constraint`. Together with **magrittr** pipes you can build a model just like a **dplyr** statement incrementally, without worrying on how to build the matrix and vectors. Furthermore, an **ompr** model is independent from specific solvers and a lot of popular solvers can easily be used through the **ROI** family of packages (Hornik et al. 2016).

The idea to model mixed-integer programs algebraically is not new in general. Domain specific languages such as GNU MathProg or the JuMP project (Dunning, Huchette, and Lubin 2015) in *Julia* implement a similar approach as **ompr** and inspired its development. As far as I know, there is one other related *R* package, **roml** (Vana, Schwendinger, and Hochreiter 2016), that is currently under development and follows a similiar pathway.

The **ompr** package is developed and available on GitHub. In addition to the package itself several vignettes and examples exist describing how to model and solve popular optimization problems, such as the traveling salesman problem, the warehouse location problem or solving Sudokus interactively with **shiny**.

In this talk I will present the modelling features of **ompr**, how the package can be used to solve practical optimization problems and some ideas for future developments.

References

Dunning, Iain, Joey Huchette, and Miles Lubin. 2015. "JuMP: A Modeling Language for Mathematical Optimization." *arXiv:1508.01982 [Math.OC]*. http://arxiv.org/abs/1508.01982.

GNU Linear Programming Kit. 2017. http://www.gnu.org/software/glpk/glpk.html.

Hornik, Kurt, David Meyer, Florian Schwendinger, and Stefan Theussl. 2016. *ROI: R Optimization Infrastructure*. https://CRAN.R-project.org/package=ROI.

Vana, Laura, Florian Schwendinger, and Ronald Hochreiter. 2016. *R Optimization Modeling Language*. https://r-forge.r-project.org/projects/roml/.

Easy imputation with the simputation package

Mark van der Loo[1]

1. *Statistics Netherlands*

Keywords: imputation, missing values, grammar of data manipulation

Webpages: github | CRAN | personal blog

Missing value imputation is a common technique for dealing with missing data. Accordingly, R and its many extension packages offer a wide range of techniques to impute missing data. Imputation can be done using specialized imputation functions or, with a bit of programming, one of the many predictive models available in R or its extension packages.

The current set of available imputation and modeling techniques is the result of decades of development by many different contributors. As a result, imputation and modeling functions may have very different interfaces accross packages. Combining and comparing imputation methods can therefore be a cumbersome task.

The **simputation** package offers a uniform and robust interface to a number of popular imputation techniques. The package follows the 'grammar of data manipulation' (Wickham and Francois 2016), where the first argument to a function and its output are always rectangular datasets. This allows one to chain imputaton methods with the not-a-pipe operator of the **magrittr** package (Bache and Wickham 2014). In **simputation** all imputation functions are of the following form.

```
impute_[model](data, formula, ...)
```

For example, functions `impute_lm` or `impute_em` impute missing values based on linear modeling or EM-estimation respectively. The formula object is interpreted so multiple variables can be imputed based on the same set of predictors. Also, a grouping operator (`|`) allows one to impute using the split-apply-combine strategy for any imputation method.

Currently supported methods include imputation based on standard linear models, M-estimation and elasticnet (ridge, lasso) regression; CART and randomForest models; multivariate methods including EM-estimation and iterative randomForest estimation; donor imputation including random and sequential hotdeck, predictive mean matching and kNN imputation. A flexible interface for simple user-provided imputation expressions is provided as well.

References

Bache, Stefan Milton, and Hadley Wickham. 2014. *Magrittr: A Forward-Pipe Operator for R.* https://CRAN.R-project.org/package=magrittr.

van der Loo, Mark. 2016. *Simputation: Simple Imputation.* https://github.com/markvanderloo/simputation.

Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

useR! 2017 Oral Presentation

Using the alphabetr package to determine paired T cell receptor sequences

Edward S. Lee[1]

1. Institute of Infection, Immunity & Inflammation, Glasgow Biomedical Research Centre, University of Glasgow

Keywords: bioinformatics, sequencing, immunology, T cell receptors

Webpages: https://cran.r-project.org/web/packages/alphabetr/index.html, https://github.com/edwardslee/alphabetr

The immune system has the monumental challenge of being capable of respond- ing to any pathogen or foreign substance invading the body while ignoring self and innocuous molecules. T cells—which play a central role in directing immune responses, regulating other immune cells, and remembering past infections— accomplish this feat by maintaining a diverse repertoire of T cell receptors (TCR). A typical T cell expresses one unique TCR, and the TCR is made up of two chains—the TCRα and TCRβ chains—that both determine the set of molecules that the T cell can respond to. Since T cells play such a central role in many immune responses, identifying the TCR pairs of T cells involved in infec- tious diseases, cancers, and autoimmune diseases can have profound insights for designing vaccines and immunotherapies. I introduce a novel approach to obtaining paired TCR sequences with the alphabetr package, which implements algorithms that identify TCR pairs in an efficient, high-throughput fashion for antigen-specific T cell populations (Lee et al. 2017).

References

Lee, Edward S, Paul G Thomas, Jeff E Mold, and Andrew J Yates. 2017. “Identifying T Cell Receptors from High-Throughput Sequencing: Dealing with Promiscuity in TCRα and TCRβ Pairing.” PLOS Computational Biology 13 (1). Public Library of Science: e1005313.

Scalable, Spatiotemporal Tidy Arrays for R (stars)

Edzer Pebesma, Etienne Racine, Michael Sumner

Spatiotemporal data often comes in the form of dense arrays, with space and time being array dimensions. Examples include socio-economic or demographic data, environmental variables monitored at fixed stations, time series of satellite images with multiple spectral bands, spatial simulations, climate model results. Currently, R does not have infrastructure to handle and analyse such arrays easily. Package raster is probably still the most powerful package for handling this kind of data in memory and on disk, but does not address non-raster time series, rasters time series with multiple attributes, rasters with mixed type attributes, or spatially distributed sets of satellite images. This project will not only deal with these cases, but also extend the "in memory or on disk" model to that where the data are held remotely in cloud storage, which is a more feasible option e.g. for satellite data collected Today. We will implement pipe-based workflows that are developed and tested on samples before they are evaluated for complete datasets, and discuss the challenges of visualiasation and storage in such workflows. This is work in progress, and the talk will discuss the design stage and hopefully show an early prototype.

Interfacing Google's spherical geometry library (S2) for spatial data in R

Ege Rubak[1]

1. Aalborg University, Denmark

Keywords: Spatial statistics, spherical geometry, geospatial index, GIS

Webpages: https://github.com/spatstat/s2, https://cran.r-project.org/package=s2

Google's S2 geometry library is a somewhat hidden gem which hasn't received the attention it deserves. It both facilitates geometric operations directly on the sphere such as polygonal unions, intersections, differences etc. without the hassle of projecting data in the common latitude and longitude format, and provides an efficient quadtree type hierarchical geospatial index.

The original *C++* source code is available in a Google Code archive and it has been partially ported to e.g. Java, Python, NodeJS, and Go, and it is used in MongoDB's 2dsphere index.

The geospatial index in the S2 library allows for useful approximations of arbitrary regions on the sphere which can be efficiently manipulated.

We describe how the geospatial index is constructed and some of it properties as well as how to perform some of the geometrical operations supported by the library. This is all done using **Rcpp** to interface the *C++* code from *R*.

ICAOD: An R Package to Find Optimal Designs for Nonlinear Models with Imperialist Competitive Algorithm

Ehsan Masoudi

Institute of Psychology, University of Münster, Germany

Keywords: Optimal design, Nonlinear models, Optimization, Evolutionary algorithm

Webpages: https://cran.r-project.org/web/packages/ICAOD

The **ICAOD** package applies a novel multi-heuristic algorithm called imperialist competitive algorithm (ICA) to find different types of optimal designs for nonlinear models (Masoudi et al., in press). The setup assumes that we have a general parametric regression model and a design criterion formulated as a convex function of the Fisher information matrix. The package constructs locally D-optimal, minimax D-optimal, standardized maximin D-optimal and optimum-on-the-average designs for a class of nonlinear models, including multiple-objective optimal designs for the 4-parameter Hill model commonly used in dose response studies and other applied fields. Several useful functions are also provided in the package, namely a function to check optimality of the generated design using an equivalence theorem followed by a graphic plot of the sensitivity function for visual appreciation. Another function is to compute the efficiency lower bound of the generated design if the algorithm is terminated prematurely.

References

Masoudi E., Holling H., Wong W.K. (in press) Application of imperialist competitive algorithm to find minimax and standardized maximin optimal designs, *Computational Statistics & Data Analysis*.

Rc^2: an Environment for Running R and Spark in Docker Containers

E. James Harner[1*] **and Mark Lilback**[2]

1. West Virginia University
2. Rc^{2}ai
*Contact author: jharner@stat.wvu.edu

Keywords: R, Spark, Docker containers, Kubernetes, Cloud computing

Rc^2 (R cloud computing) is a containerized environment for running *R*, Hadoop, and Spark with various persistent data stores including PostgreSQL, HDFS, HBase, Hive, etc. At this time, the server side of Rc^2 runs on Docker's Community Edition, which can be: on the same machine as the client, on a server, or in the cloud. Currently, Rc^2 supports a macOS client, but iOS and web clients are in active development.

The clients are designed for small or large screens with a left editor panel and a right console/output panel. The editor panel supports R scripts, R Markdown, and Sweave, but bash, SQL, Python, and additional languages will be added. The right panel allows toggling among the console and graphical objects as well as among generated help, html, and pdf files. A slide-out panel allows toggling among session files, R environments, and R packages. Extensive search capabilities are available in all panels.

The base server configuration has containers for an app server, a database server, and a compute engine. The app server communicates with the client. The compute engine is available with or without Hadoop/Spark. Additional containers can be added or removed from within Rc^2 as it is running, or various prebuilt topologies can be launched from the Welcome window. Multiple sessions can be run concurrently in tabs. For example, a local session could be running along with another session connected to a Spark cluster.

Although the Rc^2 architecture supports physical servers and clusters, the direction of computing is in virtualization. The docker containers in Rc^2 can be orchestrated by kubernetes to build arbitrarily large virtual clusters for the compute engine (e.g., parallel R) and/or for Hadoop/ Spark. The focus initially is on building a virtual cluster from Spark containers using kubernetes built on a persistent data store, e.g., HDFS. The ultimate goal is to built data science workflows, e.g., ingesting streaming data into Kafka, modulating it into a data store, and passing it to Spark Streaming.

useR! 2017 Oral Presentation

Ranking items scalably with the Bradley-Terry model

Ella Kaye[1] *and David Firth*[1,2]
1. University of Warwick, Coventry, UK
2. Alan Turing Institute, London, UK

Keywords: Citation data, Directed network, Paired comparisons, Quasi-symmetry, Sparse matrices

Webpage: https://github.com/EllaKaye/BradleyTerryScalable

Motivated by the analysis of large-scale citation networks, we implement the familiar Bradley-Terry model (Zermelo 1929; Bradley and Terry 1952) in such a way that it can be applied, with relatively modest memory and execution-time requirements, to pair-comparison data from networks with large numbers of nodes. This provides a statistically principled method of ranking a large number of objects, based only on paired comparisons.

The **BradleyTerryScalable** package complements the existing *CRAN* package **BradleyTerry2** (Firth and Turner 2012) by permitting a much larger number of objects to be compared. In contrast to **BradleyTerry2**, the new **BradleyTerryScalable** package implements only the simplest, 'unstructured' version of the Bradley-Terry model. The new package leverages functionality in the additional *R* packages **igraph** (Csardi and Nepusz 2006), **Matrix** (Bates and Maechler 2017) and **Rcpp** (Eddelbuettel 2013) to provide flexibility in model specification (whole-network versus disconnected cliques) as well as memory efficiency and speed. The Bayesian approach of Caron and Doucet (2012) is provided as an optional alternative to maximum likelihood, in order to allow whole-network ranking even when the network of paired comparisons is not fully connected.

The **BradleyTerryScalable** package can readily handle data from directed networks with many thousands of nodes. The use of the Bradley-Terry model to produce a ranking from citation data was originally advocated in Stigler (1994), and was studied in detail more recently in Varin, Cattelan, and Firth (2016); here we will illustrate its use with a large-scale network of inter-company patent citations.

References

Bates, Douglas, and Martin Maechler. 2017. "Matrix: Sparse and Dense Matrix Classes and Methods." *R Package Version 1.2-8*. http://cran.r-project.org/package=Matrix.

Bradley, Ralph Allan, and Milton E Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. the Method of Paired Comparisons." *Biometrika* 39: 324–45.

Caron, François, and Arnaud Doucet. 2012. "Efficient Bayesian Inference for Generalized Bradley–Terry Models." *Journal of Computational and Graphical Statistics* 21: 174–96.

Csardi, Gabor, and Tamas Nepusz. 2006. "The igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. http://igraph.org.

Eddelbuettel, Dirk. 2013. *Seamless R and C++ Integration with Rcpp*. New York: Springer.

Firth, David, and Heather L Turner. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48 (9). http://www.jstatsoft.org/v48/i09.

Stigler, Stephen M. 1994. "Citation Patterns in the Journals of Statistics and Probability." *Statistical Science*, 94–108.

Varin, Cristiano, Manuela Cattelan, and David Firth. 2016. "Statistical Modelling of Citation Exchange Between Statistics Journals." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179: 1–63.

Zermelo, Ernst. 1929. "Die Berechnung Der Turnier-Ergebnisse Als Ein Maximumproblem Der Wahrscheinlichkeitsrechnung." *Mathematische Zeitschrift* 29: 436–60.

Extracting Meaningful Noisy Biclusters from a Binary Big-Data Matrix using the BiBitR R Package

Ewoud De Troyer[1]*, Ziv Shkedy*[1]*, Adetayo Kasim*[2] *and Javier Cabrera*[3]
1. Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium.
2. Wolfson Research Institute, Durham University Queen's Campus, University Boulevard, Thornaby, Stockton on Tees, TS17 6BH, UK
3. Rutgers University, Busch Campus, 501 Hill Center, 110 Frelinghuysen Rd, Piscataway, NJ 08854-8019, US

Keywords: R, package, biclustering, binary data

Webpages: https://cran.r-project.org/package=BiBitR, https://github.com/ewouddt/BiBitR

Biclustering is a data analysis method that can be used to cluster the rows and columns in a (big) data matrix simultaneously in order to identify local submatrices of interest, i.e., local patterns in a big data matrix. For binary data matrices, the local submatrices that biclustering methods can identify consists of rectangles of 1's. Several methods were developed for biclustering of binary data, such as the *Bimax* algorithm proposed by Prelić et al. (2006) and the *BiBit* algorithm by Rodriguez-Baena, Perez-Pulido, and Aguilar-Ruiz (2011). However, these methods are capable to discover only perfect biclusters which means that noise is not allowed (i.e., zeros are not included in the bicluster). We present an extension for the *BiBit* algorithm (*E-BiBit*) that allows for noisy biclusters. While this method works very fast, its downside is that it often produces a large number of biclusters (typically >10000) which makes it very difficult to recover any meaningful patterns and to interpret the results. Furthermore many of these biclusters are highly overlapping.

We propose a data analysis workflow to extract meaningful noisy biclusters from binary data using an extended and 'pattern-guided' version of *BiBit* and combine it with traditional clustering/networking methods. The proposed algorithm and the data analysis workflow are illustrated using the **BiBitR** R package to extract and visualize these results.

The proposed method/data analysis flow is applied to high dimensional real life health data which contains information of disease symptoms of hundreds thousands of patients. The *E-BiBit* algorithm is used to identify homogeneous subsets of patients who share the same disease symptom profiles.

The *E-BiBit* has also been included in the **BiclustGUI** R package (De Troyer and Otava (2016), De Troyer et al. (2016)), an ensemble GUI package in which multiple biclustering and visualisation methods are implemented.

References

De Troyer, E., and M. Otava. 2016. *Package 'Rcmdrplugin.BiclustGUI': 'Rcmdr' Plug-in Gui for Biclustering.* https://ewouddt.github.io/RcmdrPlugin.BiclustGUI/aboutbiclustgui/.

De Troyer, E., M. Otava, J. D. Zhang, S. Pramana, T. Khamiakova, S. Kaiser, M. Sill, et al. 2016. "Applied Biclustering Methods for Big and High-Dimensional Data Using R." In, edited by A. Kasim, Z. Shkedy, S. Kaiser, S. Hochreiter, and W. Talloen. CRC Press Taylor & Francis Group, Chapman & Hall/CRC Biostatistics Series.

Prelić, A., S. Bleuler, P. Zimmermann, Wille A., P. Bühlmann, W. Gruissem, L. Henning, L. Thiele, and E. Zitzler. 2006. "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data." *Bioinformatics* 22: 1122–9.

Rodriguez-Baena, Domingo S., Antona J. Perez-Pulido, and Jesus S. Aguilar-Ruiz. 2011. "A Biclustering Algorithm for Extracting Bit-Patterns from Binary Dataets." *Bioinformatics* 27 (19).

Interactive and Reproducible Research for RNA Sequencing Analysis

Federico Marini[1,2] *and Harald Binder*[1]

1. *Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Mainz, 55131, Germany*
2. *Center for Thrombosis and Hemostasis (CTH), University Medical Center, Mainz, 55131, Germany*

Keywords: RNA-Seq, Exploratory Data Analysis, Differential Expression, Interactivity, Reproducibility

Webpages: http://bioconductor.org/packages/pcaExplorer/, https://github.com/federicomarini/ideal

Next generation sequencing technologies, such as RNA-Seq, generate tens of millions of reads to define the expression levels of the features of interest. A wide number and variety of software packages have been developed for accommodating the needs of the researcher, mostly in the R/Bioconductor framework. Many of them focus on the identification of differentially expressed (DE) genes (**DESeq2**, **edgeR**, (Love et al. 2015)) to discover quantitative changes between experimental groups, while other address alternative splicing, discovery of novel transcripts, or RNA editing.

Moreover, Exploratory Data Analysis is a common step to all these workflows, and despite its importance for generating highly reliable results, it is often neglected, as many of the steps involved might require a considerable proficiency of the user in the programming languages. Principal Components Analysis (PCA) is used often to obtain a dimension-reduced overview of the data (Jolliffe 2002).

Our proposal will address the two steps of Exploratory Data Analysis and Differential Expression analysis with two different packages, integrated and available in Bioconductor, namely **pcaExplorer** and **ideal**. We propose web applications developed in the Shiny framework which will also include support for reproducible analyses, thanks to an embedded text editor and a template document, to seamlessly generate HTML reports as a result of the user's exploration.

This solution, which we also outlined in (Marini and Binder 2016), serves as a concrete proof of principle of integrating the essential features of interactivity (as a proxy for accessibility) and reproducibility in the same tool, fitting both the needs of life scientists and experienced analyists, thus making our packages good candidates to become companion tools for each RNA-Seq analysis.

References

Jolliffe, I T. 2002. "Principal Component Analysis, Second Edition." *Encyclopedia of Statistics in Behavioral Science* 30 (3): 487. doi:10.2307/1270093.

Love, Michael I., Simon Anders, Vladislav Kim, and Wolfgang Huber. 2015. "RNA-Seq workflow: gene-level exploratory analysis and differential expression." *F1000Research* 4: 1070. doi:10.12688/f1000research.7035.1.

Marini, Federico, and Harald Binder. 2016. "Development of Applications for Interactive and Reproducible Research : a Case Study." *Genomics and Computational Biology* 3 (1): 1–4.

Data Carpentry: Teaching Reproducible Data Driven Discovery

François Michonneau[1] *and Tracy Teal*[2]

1. Whitney Laboratory for Marine Bioscience, University of Florida
2. Data Carpentry

Keywords: reproducible research, rmarkdown, open science, training, github

Webpages: https://datacarpentry.org

Data Carpentry is a non-profit organization and community. It develops and teaches workshops aimed at researchers with little to no programming experience. It teaches skills and good practices for data management and analysis, with a particular emphasis on reproducibility. Over a two-day workshop, participants are exposed to the full life cycle of data-driven research. Since its creation in 2014, Data Carpentry has taught over 125 workshops and trained 400+ certified instructors. Because the workshops are domain specific, participants can get familiar with the dataset used throughout the workshop quickly, and focus on learning the computing skills. We have developed detailed assessments to evaluate the effectiveness and level of satistaction of the participants after attending a workshop as well as the impact on their research and careers 6 months or more after a workshop. Here, we will present an overview of the organization, the skills taught with a particular emphasis on using R, and the strategies used to make these workshops successful.

Stream processing with R in AWS

Gergely Daroczi[1]

1. CARD.com

Keywords: stream processing, big data, ETL, scale

Webpages: https://CRAN.R-project.org/package=AWR, https://CRAN.R-project.org/package=AWR.KMS, https://CRAN.R-project.org/package=AWR.Kinesis

R is rarely mentioned among the big data tools, although it's fairly well scalable for most data science problems and ETL tasks. This talk presents an open-source *R* package to interact with Amazon Kinesis via the MultiLangDaemon bundled with the Amazon KCL to start multiple *R* sessions on a machine or cluster of nodes to process data from theoretically any number of Kinesis shards.

Besides the technical background and a quick introduction on how Kinesis works, this talk will feature some stream processing use-cases at CARD.com, and will also provide an overview and hands-on demos on the related data infrastructure built on the top of Docker, Amazon ECS, ECR, KMS, Redshift and a bunch of third-party APIs – besides the related open-source *R* packages, eg **AWR**, **AWR.KMS** and **AWR.Kinesis**, developed at CARD.

References

- AWS SDK for Java
- Amazon Kinesis Client Library for Java
- Amazon Kinesis Dev Guide

The analysis of R learning styles with R

Gert Janssenswillen[1,2] and Benoît Depaire[1] | 1. UHasselt - Hasselt University, Faculty of Business Informatics, Agoralaan , 3590 Diepenbeek, Belgium | 2. FWO - Research Foundation Flanders, Egmontstraat 6, 1000 Brussels, Belgium

Keywords: R in education, Learning patterns, Learning styles

The world of education is changing more than ever. In the university of the 21st century, there is no room for one-way education with a summative evaluation at the end of the teaching period. Instead, there is a need for formative assessment, including frequent and individual feedback (Lindblom Ylanne and Lonka 1998). However, when the number of students is large, providing individual feedback requires a huge amount of effort and time. This effort is intensified when the subject matter taught allows for a certain flexibility to solve problems. Although data analysis can provide useful insights about learning styles and patterns of students both at the time of learning and afterwards, this abstract shows that it can also be leveraged for providing fast feedback.

This abstract both incorporates a procedure to provide large-scale (semi-)individual feedback by using systematic assignments, as well as insights into different learning styles by combining information on the assigments and the final scores of the students. The case used is a course on explorative data analysis (EDA) taught to a group of circa 80 first year business engineering students at Hasselt University, covering a diverse set of topics such as data manipulation, visualization, import and tidying. During the course, students have to complete assignments on a regular interval in order to fully administer the new skills, each arranged around a specific topic. These assignments come in the form of Rmarkdown files in which the students have to complete R-chunks appropriately. Each Rmarkdown file is then re-run by the education team, and the data generated for each student is used for evaluation.

Each problem which the students have to solve in these assignments is labelled by the education team with the principles it assesses. For example, in the case of visualization, it might have to do with using appropriate aestethics, appropriate geoms, appropriate context (e.g. titles, labels), etc. By mapping these labels and the scores of the student, a precise *learning profile* for each student can be constructed which indicates his weaknesses and his strenghts (Vermunt and Vermetten 2004). By using this information, students can be clustered in different groups, which can then be addressed with tailored feedback on their progress and pointers to useful additional exercises in order to remedy those areas in which they perform less good.

In a second step, an ex post analysis can be done by combining the learning profiles created with the final grades and possibly other information such as educational background. This information can be employed to find which group of students represent *problem cases*, i.e. having a high probability of failing for the course. These insights can proof useful in future editions of the course, as a mechanism for rapid identification of students who might have difficulties with certain concepts. Moreover, it can be used to adapt the course, such that certain concepts which proof the be problematic are highlighted in a different or in a more comprehensive manner throughout the course (Tait and Entwistle 1996).

References

Lindblom Ylanne, Sari, and Kirsti Lonka. 1998. "Individual Ways of Interacting with the Learning Environment. Are They Related to Study Success?" *Learning and Instruction* 9 (1). Elsevier: 1–18.

Tait, Hilary, and Noel Entwistle. 1996. "Identifying Students at Risk Through Ineffective Study Strategies." *Higher Education* 31 (1). Springer: 97–116.

Vermunt, Jan D, and Yvonne J Vermetten. 2004. "Patterns in Student Learning: Relationships Between Learning Strategies, Conceptions of Learning, and Learning Orientations." *Educational Psychology Review* 16 (4). Springer: 359–84.

useR! 2017 Oral Presentation

Too good for your own good: Shiny prototypes out of control

Grace Meyer

1. *Mango Solutions*
2. *RLadies London*

Keywords: Shiny, Project Management, Product Management, Best Practice, Stakeholder Management

Shiny development is exploding in the *R* world, especially for enabling analysts to share their results with business users interactively. At Mango Solutions, the number of **Shiny** apps we are being commissioned to build has increased dramatically with approximately 30% of current projects involving some aspect of **Shiny** development.

Typically, **Shiny** has been used as a prototyping tool to quickly show business the value of data driven projects with the aim to productionalise the app once buy-in from stakeholders is gained. **Shiny** is fantastically quick to get an app up and running and into the hands of users and additional features can be rapidly prototyped for stakeholders.

In this presentation I will share with you our experience from a client project where ***Shiny*** *prototyping got out of control*- the app was so successful for the business the pilot phase quickly evolved into full deployment as more users were involved in "testing" without production best practice implemented yet. I will then tell you how we faced into this challenge which involved client education and the planning and implementation of the required deployment rigour.

I will also share our thoughts on how to approach **Shiny** prototyping and development (taking on board our lessons learnt) depending on the app's needs- you can still quickly implement features with **Shiny** but with a few recommendations you can minimise the largest risks of your app getting away from you.

Daff: diff, patch and merge for data.frames

Gregory R. Warnes[1] *and Edwin de Jonge*[2]

1. Medidata Solutions Inc. (https://www.mdsol.com/)
2. Statistics Netherlands (https://www.cbs.nl/)

Keywords: Reproducible research, data versioning

Webpages: https://CRAN.R-project.org/package=daff, https://github.com/edwindj/daff

In data analysis, it can be necessary to compare two files comparing tabular data. Unfortunately, existing tools have been customized for comparing source code or other text files, and are unsuitable for comparing tabular data.

The *daff* R package provides tools for comparing and tracking changes in tabular data stored in data.frames. *daff* wraps Paul Fitz's multi-language *daff* package (https://github.com/paulfitz/daff), which generates data diff that capture row and column modifications, reorders, additions, and deletions. These data diffs follow a standard format (http://dataprotocols.org/tabular-diff-format/) which can be used to HTML formatted diffs, summarize changes, and even *patch* (a new version of) input data.

daff augments brings the utility of source-code change tracking tools to tabular data, enabling data versioning as a component of software development and reproducible research.

References

Fitzpatrick, Paul. 2014. "Coopy Highlighter Diff Format for Tables," May. http://specs.okfnlabs.org/tabular-diff-format.

R goes Mobile: Efficient Scheduling for Parallel R Programs on Heterogeneous Embedded Systems

Helena Kotthaus, Andreas Lang, Olaf Neugebauer, Peter Marwedel

TU Dortmund University, Department of Computer Science 12

Keywords: Parallelization, Resource-Aware Scheduling, Hyperparameter Tuning, Embedded Systems

Webpages: http://sfb876.tu-dortmund.de/SPP/sfb876-a3.html

We present a resource-aware scheduling strategy for parallelizing *R* applications on heterogeneous architectures, like those commonly found in mobile devices. Such devices typically consist of different processors with different frequencies and memory sizes, and are characterized by tight resource and energy restrictions. Similar to the **parallel** package that is part of the *R* distribution, we target problems that can be decomposed into independent tasks that are then processed in parallel. However, as the **parallel** package is not resource-aware and does not support heterogeneous architectures, it is ill-suited for the kinds of systems we are considering.

The application we are focusing on is parameter tuning of machine learning algorithms. In this scenario, the execution time of an evaluation of a parameter configuration can vary heavily depending on the configuration and the underlying architecture. Key to our approach is a regression model that estimates the execution time of a task for each available processor type based on previous evaluations. In combination with a scheduler allowing to allocate tasks to specific processors, we thus enable efficient resource-aware parallel scheduling to optimize the overall execution time.

We demonstrate the effectiveness of our approach in a series of examples targeting the ARM big.LITTLE architecture, an architecture commonly found in mobile phones.

References

ARM. 2017. "big.LITTLE Technology." https://www.arm.com/products/processors/technologies/biglittleprocessing.php.

Helena Kotthaus, Ingo Korb. 2017. "TraceR: Profiling Tool for the R Language." Department of Computer Science 12, TU Dortmund University. https://github.com/allr/traceR-installer.

Kotthaus, Helena, Ingo Korb, and Peter Marwedel. 2015. "Performance Analysis for Parallel R Programs: Towards Efficient Resource Utilization." 01/2015. Department of Computer Science 12, TU Dortmund University.

Richter, Jakob, Helena Kotthaus, Bernd Bischl, Peter Marwedel, Jörg Rahnenführer, and Michel Lang. 2016. "Faster Model-Based Optimization Through Resource-Aware Scheduling Strategies." In *LION10*, 267–73. Springer International Publishing.

Developing and deploying large scale Shiny applications for non-life insurance

Author Herman Sontrop[1]

1. FRISS | fraud, risk & compliance. Utrecht, the Netherlands.

Keywords: Shiny modules, HTMLWidgets, HTMLTemplates, openCPU, NoSQL, Docker

Webpages: https://www.friss.eu/en/

FRISS is a Dutch, fast growing company with a 100% focus on fraud, risk and compliance for non-life insurance companies and is the European market leader with over 100+ implementations in more than 15 countries worldwide. The FRISS platform offers insurers fully automated access to a vast set of external data sources, which together facilitate many different types of screenings, based on knowledge rules, statistical models, clustering, text mining, image recognition and other machine learning techniques. The information produced by the FRISS platform is bundled into a risk score that provides a quantified risk assessment on a person or case, that enables insurers to make better and faster decisions.

At FRISS, all analytical applications and services are based on *R*. Interactive applications are based on *Shiny*, a popular web application platform for *R* designed by RSTUDIO, while *openCPU*, an interoperable HTTP API for *R*, is used to deploy advanced scoring engines at scale, that can be deeply integrated into other services.

In this talk, we show various architectures on how to create high performance, large scale *Shiny* apps and scoring engines, with a clean code base. *Shiny* apps are based around the *module pattern*, *HTMLWidgets* and *HTMLTemplates*. Shiny modules allow a developer to compose a complex app via a set of easy to understand modules, each with separate UI and server logic. In these architectures, each module has a set of reactive inputs and outputs and focuses on a single, dedicated task. Subsequently, the modules are combined in a main app that can perform a multitude of complex tasks, yet is still easy to understand and to reason about. In addition, we show how *HTMLWidgets* allow you to bring the best of *JavaScript*, the language of the web, into *R* and show how *HTMLTemplates* can be used to create *R* based web applications with a fresh, modern and distinct look.

Finally, in this talk, we show various real-life examples of complex, large scale *Shiny* applications developed at FRISS. These applications are actively used by insurers worldwide for reporting, dashboarding, anomaly detection, interactive network exploration and fraud detection and allow insurers to combat fraud, risk and compliance. In addition, we show how the aforementioned techniques can be combined with modern NoSQL databases like *ElasticSearch*, *MongoDB* and *Neo4j*, to create high performance apps and how *Docker* can be used for a smooth deployment process in on-premises scenarios, that is both fast and secure.

Quantitative fisheries advice using R and FLR

Iago Mosqueira[1], Ernesto Jardim[1], Finlay Scott[1], Laurence T. Kell[2]

1. European Commission, DG Joint Research Centre, Directorate D - Sustainable Resources, Unit D.02 Water and Marine Resources, Via E. Fermi 2749, 21027 Ispra VA, Italy.

2. International Commission for the Conservation of Atlantic Tuna (ICCAT) Secretariat - Corazón de María, 8. 28002 Madrid, SPAIN

Keywords: Quantitative Fisheries Science, Common Fisheries Policy, Management Strategy Evaluation, advice, simulation

Webpages: https://flr-project.org, https://github.com/flr

The management of the activities of fishing fleets aims at ensuring the sustainable exploitation of the ocean's living resources, the provision of important food resources to humankind, and the profitability of an industry that is an important economic and social activity in many areas of Europe and elsewhere. These are the principles of the European Union Common Fisheries Policy (CFP), which has driven the management of Europe's fisheries resources since 1983.

Quantitative scientific advice is at the heart of fisheries management regulations, providing estimates of the likely current and future status of fish stocks through statistical population models, termed stock assessments, but also probabilistic comparisons of the expected effects of alternative management procedures. Management Strategy Evaluation (MSE) uses stochastic simulation to incorporate both the inherent variability of natural systems, and our limited ability to model their dynamics, into analyses of the expected effects of a given management intervention on the sustainability of both fish stocks and fleets.

The Fishery Library in R (*FLR*) project has been for the last ten years building an extensible toolset of statistical and simulation methods for quantitative fisheries science (Kell et al. 2007), with the overarching objective of enabling fisheries scientists to carry out analyses of management procedures in a simplified and robust manner through the MSE approach.

FLR has become widely used in many of the scientific bodies providing fisheries management advice, both in Europe and elsewhere. The evaluation of the effects of some elements of the revised CFP, the analysis of the proposed fisheries management plans for the North Sea, or the comparison of management strategies for Atlantic tuna stocks, among others, have used the *FLR* tools to advice managers of the possible courses of action to favour the sustainable use of many marine fish stocks.

The *FLR* toolset is currently composed of 20 packages, covering the various steps in the fisheries advice and simulation workflow. They include a large number of S4 classes, and more recently Reference Classes, to model the data structures that represent each of the elements in the fisheries system. Class inheritance and method overloading are essential tools that have allowed the *FLR* packages to interact, complement and enrich each other, while still limiting the number of functions an user needs to be aware of. Methods also exist that make use of R's parallelization facilities and of compiled code to deal with complex computations. Statistical models have also been implemented, making use of both *R*'s capabilities and external libraries for Automatic Differentiation.

We present the current status of *FLR*, the new developments taking place, and the challenges faced in the development of a collection of packages based on S4 classes and methods.

References

Kell, L. T., I. Mosqueira, P. Grosjean, J.-M. Fromentin, D. Garcia, R. Hillary, E. Jardim, et al. 2007. “FLR: An Open-Source Framework for the Evaluation and Development of Management Strategies.” *ICES Journal of Marine Science* 64 (4). http://dx.doi.org/10.1093/icesjms/fsm012.

implyr: A **dplyr** Backend for a Apache Impala

Ian Cook[1]

1. Cloudera

Keywords: Tidyverse, **dplyr**, SQL, Apache Impala, Big Data

Webpages: https://CRAN.R-project.org/package=implyr

This talk introduces **implyr**, a new **dplyr** backend for Apache Impala (incubating). I compare the features and performance of **implyr** to that of **dplyr** backends for other distributed query engines including **sparklyr** for Apache Spark's Spark SQL, **bigrquery** for Google BigQuery, and **RPresto** for Presto.

Impala is a massively parallel processing query engine that enables low-latency SQL queries on data stored in the Hadoop Distributed File System (HDFS), Apache HBase, Apache Kudu, and Amazon Simple Storage Service (S3). The distributed architecture of Impala enables fast interactive queries on petabyte-scale data, but it imposes limitations on the **dplyr** interface. For example, row ordering of a result set must be performed in the final phase of query processing. I describe the methods used to work around this and other limitations.

Finally, I discuss broader issues regarding the **DBI**-compatible interfaces that **dplyr** requires for underlying connectivity to database sources. **implyr** is designed to work with any **DBI**-compatible interface to Impala, such as the general packages **odbc** and **RJDBC**, whereas other **dplyr** database backends typically rely on one particular package or mode of connectivity.

mlrHyperopt: Effortless and collaborative hyperparameter optimization experiments

Jakob Richter[1] and Jörg Rahnenführer[1] and Michel Lang[1]

1. TU Dortmund University, Germany

Keywords: machine learning, hyperparameter optimization, tuning, classification, networked science

Webpages: https://jakob-r.github.io/mlrHyperopt/

Most machine learning tasks demand hyperparameter tuning to achieve a good performance. For example, Support Vector Machines with radial basis functions are very sensitive to the choice of both kernel width and soft margin penalty C. However, for a wide range of machine learning algorithms these "search spaces" are less known. Even worse, experts for the particular methods might have conflicting views. The popular package **caret** (Jed Wing et al. 2016) approaches this problem by providing two simple optimizers *grid search* and *random search* and individual search spaces for all implemented methods. To prevent training on misconfigured methods a *grid search* is performed by default. Unfortunately it is only documented which parameters will be tuned but the exact bounds have to be obtained from the source code. As a counterpart **mlr** (Bischl et al. 2016) offers more flexible parameter tuning methods such as an interface to **mlrMBO** (Bischl et al. 2017) for conducting Bayesian optimization. Unfortunately **mlr** lacks of default search spaces and thus parameter tuning becomes difficult. Here **mlrHyperopt** steps in to make hyperparameter optimization as easy as in **caret**. As a matter of fact, for a developer of a machine learning package, it is unquestionable impossible to be an expert of all implemented methods and provide perfect search spaces. Hence **mlrHyperopt** aims at:

- improving the search spaces of **caret** with simple tricks.
- letting the users submit and download improved search spaces to a database.
- providing advanced tuning methods interfacing **mlr** and **mlrMBO**.

A study on selected data sets and numerous popular machine learning methods compares the performance of the grid and random search implemented in **caret** to the performance of **mlrHyperopt** for different budgets.

References

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "Mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170): 1–5. https://CRAN.R-project.org/package=mlr.

Bischl, Bernd, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas, and Michel Lang. 2017. "mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions." *arXiv:1703.03373 [Stat]*, March. http://arxiv.org/abs/1703.03373.

Jed Wing, Max Kuhn. Contributions from, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2016. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

moodler: A new R package to easily fetch data from Moodle

Jakub Chromec and Libor Juhanak
Masaryk University

Keywords: Moodle, SQL, tidy data

Webpages: https://github.com/jchrom/moodler

Learning management systems (LMS) generate large amounts of data. The LMS Moodle is at the forefront of open source learning platforms, and thanks to its widespread adoption by schools and businesses, it represents a great target for educational data-analytic efforts. In order to facilitate data analysis of Moodle data in *R*, we introduce a new *R* package: **moodler**. It is a collection of useful *SQL* queries and data-wrangling functions that fetch data from Moodle database and turn it into tidy data frames. This makes it easy to feed data from Moodle to a large number of *R* packages that focus on specific types of analyses.

RQGIS - integrating *R* with QGIS for innovative geocomputing

Jannes Muenchow[1], Patrick Schratz[1] and Alexander Brenning[1]

1. Friedrich Schiller University of Jena

Keywords: GIS interface, QGIS, Python interface

Webpages: https://cran.r-project.org/web/packages/RQGIS/index.html, https://github.com/jannes-m/RQGIS

RQGIS establishes an interface to QGIS - the most widely used open-source desktop geographical information system (GIS). Since QGIS itself provides access to other GIS (SAGA, GRASS, GDAL, etc.), **RQGIS** brings more than 1000 geoalgorithms to the *R* console. Furthermore, *R* users do not have to touch *Python* though **RQGIS** makes use of the QGIS *Python* API in the background. Also, several convenience functions facilitate the usage of **RQGIS**. For instance, `open_help` provides instant access to the online help and `get_args_man` automatically collects all function arguments and respective default values of a specified geoalgorithm. The workhorse function `run_qgis` also accepts spatial objects residing in *R*'s global environment as input, and also loads QGIS output (such as shapefiles and rasters) directly back into *R*, if desired. Here, we will demonstrate the fruitful combination of *R* and QGIS by spatially predicting plant species richness of a Peruvian fog-oasis. For this, we will use **RQGIS** to extract terrain attributes (from a digital elevation model) which subsequently will serve as predictors in a non-linear Poisson regression. Apart from this, there are many more useful applications that combine *R* with GIS. For instance, GIS technologies include among others algorithms for the computation of stream networks, surface roughness, terrain classification, landform identification as well as routing and spatial neighbor operations. On the other hand, *R* provides access to advanced modeling techniques, kriging interpolation, and spatial autocorrelation and spatial cross-validation algorithms to name but a few. Naturally, this paves the way for innovative and advanced statistical geocomputing. Compared to other *R* packages integrating GIS functionalities (**rgrass7**, **RSAGA**, **RPyGeo**), **RQGIS** accesses a wider range of GIS functions, and is often easier to use. To conclude, anyone working with large spatio-temporal data in *R* may benefit from the *R*-QGIS integration.

R Package **glmm**: Likelihood-Based Inference for Generalized Linear Mixed Models

Christina Knudson, Charles Geyer, Galin Jones

Keywords: generalized linear mixed model, maximum likelihood, statistical inference, Monte Carlo

Webpages: https://CRAN.R-project.org/package=glmm, https://cknudson.com

The flexibility of generalized linear mixed models (GLMMs) allows researchers to model correlated data, perhaps with a response variable that is binomial- or Poisson-distributed. However, likelihood-based inference is limited in practice because the likelihood function for a GLMM is often a high-dimensional integral. Recent advances in statistical theory have brought about a new *R* package for modeling and analyzing GLMMs. Using intuitive examples, this session will demonstrate a few methods of likelihood-based inference made possible with *R* package **glmm**. No prior experience with GLMMs is required.

Christina Knudson will be an assistant professor at the University of St. Thomas beginning in September, 2017. Charles Geyer and Galin Jones are professors at the University of Minnesota.

References

Geyer C. (1990). *Likelihood and Exponential Families.* PhD thesis, University of Washington.

Geyer C.J. (1994). “On the Convergence of Monte Carlo Maximum Likelihood Calculations.” *Journal of the Royal Statistical Society, Series B*, 61, 261-274.

Geyer C.J., Thompson E. (1992). “Constrained Monte Carlo Maximum Likelihood for Dependent Data.” *Journal of the Royal Statistical Society, Series B*, 54, 657-699.

Knudson C. (2015). *glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation.* R package version 1.0.2, URL http://CRAN.R-project.org/package=glmm.

Knudson C. (2016). *Monte Carlo Likelihood Approximation for Generalized Linear Mixed Models.* Ph.D. Thesis, University of Minnesota.

Sung Y.J., Geyer C.J. (2007). “Monte Carlo Likelihood Inference for Missing Data Models.” *Annals of Statistics*, 35, 990-1011.

odbc - A modern database interface

Jim Hester[1]

1. RStudio Inc.

Keywords: ODBC, DBI, databases, dplyr, RStudio

Webpages: https://CRAN.R-project.org/package=odbc, https://github.com/rstats-db/odbc

Getting data into and out of databases is one of the most fundamental parts of data science. Much of the world's data is stored in databases, including traditional databases such as *SQL Server*, *MySQL*, *PostgreSQL*, and *Oracle*, as well as non-traditional databases like *Hive*, *BigQuery*, *Redshift* and *Spark*.

The **odbc** package provides an R interface to Open Database Connectivity (ODBC) drivers and databases including all those listed previously. **odbc** provides consistent output; including support for timestamps and 64-bit integers, improved performance for reading and writing, and complete compatibility with the **DBI** package.

odbc connections can be used as **dplyr** backends, allowing one to perform expensive queries within the database and reduce the need to transfer and load large amounts of data in an R session. **odbc** is also integrated into the RStudio IDE, with dialogs to setup and establish connections, preview available tables and schemas and run ad-hoc SQL queries. The RStudio Professional Products are bundled with a suite of ODBC drivers, to make it easy for System Administrators to establish and support connections to a variety of database technologies.

manifestoR - a tool for data journalists, a source for text miners and a prototype for reproducibility software

Jirka Lewandowski[1]
1. WZB Berlin Social Science Center

Keywords: political science, reproducibility, corpus, data journalism, text mining

Webpages: CRAN | package webpage

The Manifesto Project is a long-term political science research project that has been collecting, archiving and analysing party programs from democratic elections since 1979, and is one of longest standing and most widely used data sources in political science. The project recently released **manifestoR** as its official *R* package for accessing and analysing the data collected by the project. The package is aimed at three groups: it is a valuable tool for data journalism and social sciences, a data source for text mining, and a prototype for software that promotes research reproducibility.

The **manifestoR** package provides access to the Manifesto Corpus (Merz, Regel & Lewandowski 2016) – the project's text database – which contains more than 3000 digitalised election programmes from 573 parties, together running in elections between 1946 and 2015 in 50 countries, and includes documents in more than 35 different languages. More than 2000 of these documents are available as digitalised, cleaned, UTF-8 encoded full text – the rest as PDF files. As these texts are accessible from directly within *R*, **manifestoR** provides a comfortable and valuable data source for text miners interested in political and/or multilingual training data, as well as for data journalists.

The manifesto texts accessible through **manifestoR** are labelled statement by statement, according to a 56 category scheme which identifies policy issues and positions. On the basis of this labelling scheme, the political science community has developed many aggregate indices on different scales for parties' ideological positions. Most of these algorithms have been collected and included in **manifestoR** in order to provide a centralised and easy to use starting point for scientific and journalistic analyses and inquiries.

Replicability and reproducibility of scientific analyses are core values of the *R* community, and are of growing importance in the social sciences. Hence, **manifestoR** was designed with the goal of reproducible research in mind and tries to set an example of how a political science research project can publish and maintain an open source package to promote reproducibility when using its data. The Manifesto Project's text collection is constantly growing and being updated, but any version ever published can easily be used as the basis for scripts written with **manifestoR**. In addition, the package integrates seamlessly with the widely-used **tm** package (Feinerer 2008) for text mining in *R*, and provides a `data_frame` representation for every data object in order to connect to the **tidyverse** packages (Wickham 2014), including the text-specific **tidytext** (Silge & Robinson 2016). For standardising and open-sourcing the implementations of aggregate indices from the community in **manifestoR**, we sought collaboration with the original authors. Additionally, the package provides infrastructure to easily adapt such indices, or to create new ones. The talk will also discuss the lessons learned and the unmet challenges that have arisen in developing such a package specifically for the political science community.

References

- Feinerer, Ingo (2008). A text mining framework in R and its applications. Doctoral thesis, WU Vienna University of Economics and Business.
- Merz, N., Regel, S., & Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. Research & Politics, 3(2), 2053168016643346. doi: 10.1177/2053168016643346
- Silge, J., & Robinson, D. (2016). Tidytext: Text Mining and Analysis Using Tidy Data Principles in R. JOSS 1 (3). The Open Journal. doi:10.21105/joss.00037.
- Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

jamovi: a spreadsheet for R

Jonathon Love[1,2], *Damian Dropmann*[1] *and Ravi Selker*[1]

1. *the jamovi project*
2. *University of Newcastle*

Keywords: Spreadsheet, User-interface, Learning R

Webpages: https://www.jamovi.org, https://CRAN.R-project.org/package=jmv

In spite of the availability of the powerful and sophisticated *R* ecosystem, spreadsheets such as *Microsoft Excel* remain ubiquitous within the business community, and spreadsheet like software, such as *SPSS*, continue to be popular in the sciences. This likely reflects that for many people the spreadsheet paradigm is familiar and easy to grasp.

The jamovi project aims to make *R* and its ecosystem of analyses accessible to this large body of users. *jamovi* provides a familiar, attractive, interactive spreadsheet with the usual spreadsheet features: data-editing, filtering, sorting, and real-time recomputation of results. Significantly, all analyses in *jamovi* are powered by *R*, and are available from CRAN. Additionally, *jamovi* can be placed in 'syntax mode', where the underlying *R* code for each analysis is produced, allowing for a seamless transition to an interactive *R* session.

We believe that *jamovi* represents a significant opportunity for the authors of *R* packages. With some small modifications, an *R* package can be augmented to run inside of *jamovi*, allowing *R* packages to be driven by an attractive user-interface (in addition to the normal *R* environment). This makes R packages accessible to a much larger audience, and at the same time provides a clear pathway for users to migrate from a spreadsheet to *R* scripting.

This talk introduces *jamovi*, introduces its user-interface and feature set, and demonstrates the ease with which *R* packages can be augmented to additionally support the interactive spreadsheet paradigm.

jamovi is available from www.jamovi.org

Interacting with databases from Shiny

Barbara Borges Ribeiro^1

1. RStudio

Keywords: databases, shiny, DBI, dplyr, pool

Webpages: http://shiny.rstudio.com/articles/overview.html, https://github.com/rstudio/pool

Connecting to an external database from *R* can be challenging. This is made worse when you need to interact with a database from a live Shiny application. To demystify this process, I'll do two things.

First, I'll talk about best practices when connecting to a database from Shiny. There are three important packages that help you with this and I'll weave them into this part of the talk. The **DBI** package does a great job of standardizing how to establish a connection, execute safe queries using *SQL* (goodbye *SQL* injections!) and close the connection. The **dplyr** package builds on top of this to make even easier to connect to databases and extract data, since it allows users to query the database using regular **dplyr** syntax in *R* (no *SQL* knowledge necessary). Yet a third package, **pool**, exists to help you when using databases in Shiny applications, by taking care of connection management, and often resulting in better performance.

Second, I'll demo these concepts in practice by showing how we can connect to a database from Shiny to create a CRUD application. I will show the application running and point out specific parts of the code (which will be publicly available).

How the R Consortium is Supporting the R Community

Joseph Rickert[1,3] and David Smith[2,3]

1. *RStudio*
2. *Microsoft*
3. *The R Consortium*

Keywords: R Consortium

Webpages: https://www.r-consortium.org/

There is a lot happening at R Consortium! We now have 15 members, including the Gordon and Betty Moore Foundation which joined this year as a Platinum member, 21 active projects and a fired- up grant process. This March the ISC awarded grants to 10 new projects totaling nearly $240,000. In this talk we will describe how the R Consortium is evolving to carry out its mission to provide support for the R language, the R Foundation and the R Community. We will summarize the active projects, why they are important and where we think they are going, and describe how individual R users can get involved with R Consortium projects.

Text mining, the tidy way

Julia Silge [1] and David Robinson[1]

1. Stack Overflow

Keywords: text mining, natural language processing, tidy data, sentiment analysis

Webpages: https://CRAN.R-project.org/package=tidytext, http://tidytextmining.com/

Unstructured, text-heavy data sets are increasingly important in many domains, and tidy data principles and tidy tools can make text mining easier and more effective. We introduce the **tidytext** package for approaching text analysis from a tidy data perspective. We can manipulate, summarize, and visualize the characteristics of text using the *R* tidy tool ecosystem; these tools extend naturally to many text analyses and allow analysts to integrate natural language processing into effective workflows already in wide use. We explore how to implement approaches such as sentiment analysis of texts and measuring tf-idf to quantify what a document is about.

Exploring and presenting maps with **tmap**

Martijn Tennekes[1]

1. Statistics Netherlands

Keywords: Visualisation, maps, interaction, exploration

Webpages: https://CRAN.R-project.org/package=tmap, https://github.com/mtennekes/tmap

A map tells more than a thousand coordinates. Generally, people tend to like maps, because they are appealing, recognizable, and often easy to understand. Maps are not only useful for navigation, but also to explore, analyse, and present spatial data.

The **tmap** package offers a powerful engine to visualize maps, both static and interactive. It is based on the Grammar of Graphics, with a syntax similar to **ggplot2**, but tailored to spatial data. Layers from different shapes can be stacked, map legends and attributes can be added, and small multiples can be created.

An example of a map is the following. This maps consists of a choropleth of Happy Planet Index values per country and a dot map of large world cities on top. Alternatively, a choropleth can also be created with `qtm(World, "HPI")`.

```
library(tmap)
data(World, metro)

tm_shape(World) +
  tm_polygons("HPI", id = "name") +
  tm_text("name", size = "AREA") +
tm_shape(metro) +
  tm_dots(id = "name") +
tm_style_natural()
```

Interaction with charts and maps is not considered as a nice extra feature anymore, of which users will say "wow, this is interactive!". To the contrary, users will expect charts and maps to be interactive, especially when published online. Also in *R*, interaction has become common ground, especially since the introduction of **shiny** and **htmlwidgets**. However, the increase of interactive maps does not mean the end of static maps. Newspapers, journals, and posters still rely on printed maps. To design a static thematic map that is appealing, informative, and simple is a special craft.

There are two modes in which maps can be visualized: `"plot"` for static plotting and `"view"` for interactive viewing. Users are able to switch between these modes without effort. The choropleth above is reproduced in interactive mode as follows:

```
tmap_mode("view")
last_map()
```

For lazy users like me, the code `ttm()` toggles between the two modes. The created maps can be exported to static file formats, such as *pdf* and *png*, as well as interactive *html* files. Maps can also be embedded in **rmarkdown** documents and **shiny** apps.

```
save_tmap(filename = "map.png", width = 1920)
save_tmap(filename = "index.html")
```

Visualization of spatial data is important troughout the whole process from exploration to presentation. Exploration requires short and intuitive coding. Presentation requires full control over the map layout, including color scales and map attributes. The **tmap** package facilitates both exploration and presentation of spatial data.

Detecting eQTLs from high-dimensional sequencing data using recount2

Kai Kammers[1,2], Leonardo Collado-Torres[2,3,4], Margaret A Taub[2,3], and Jeffrey T Leek[2,3]

1. *Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine*
2. *Center for Computational Biology, Johns Hopkins University*
3. *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
4. *Lieber Institute for Brain Development, Johns Hopkins Medical Campus*

Keywords: eQTLs, RNA-seq, recount2, Batch Effect, gEUVADIS

Webpages: https://jhubiostatistics.shinyapps.io/recount/, https://www.bioconductor.org/packages/recount

`recount2` is a recently launched multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets for 2,041 different studies with over 70,000 human RNA-seq samples from the Sequence Read Archive (SRA), Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) projects (Collado-Torres et al. (2016)). The raw sequencing reads were processed with Rail-RNA as described at Nellore et al. (2016). RangedSummarizedExperiment objects at the gene, exon or exon-exon junctions level, the raw counts, the phenotype metadata used, the urls to the sample coverage bigWig files or the mean coverage bigWig file for a particular study can be accessed via the Bioconductor package **recount** or via a Shiny App.

We use this source of preprocessed RNA-seq expression data to present our recently developed analysis protocol for performing extensive eQTL analyses. The goal of an eQTL analysis is to detect patterns of gene expression related to specific genetic variants. We demonstrate how to integrate gene expression data from `recount2` and genotype information to perform eQTL analyses and visualize the results with gene-SNP interaction plots. We explain in detail how expression and genotype data are filtered, transformed, and batch corrected. We also discuss possible pitfalls and artifacts that may occur when analyzing genomic data from different sources jointly. Our protocol is tested on a publicly available data set of the RNA-sequencing project from the GEUVADIS consortium and also applied to recently generated omics data from the GeneSTAR project at Johns Hopkins University.

References

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey Leek. 2016. "Recount: A Large-Scale Resource of Analysis-Ready RNA-seq Expression Data." *bioRxiv.* doi:10.1101/068478.

Nellore, Abhinav, Leonardo Collado-Torres, Andrew E Jaffe, Jose Alquicira-Hernandez, Christopher Wilks, Jacob Pritt, James Morton, Jeffrey T Leek, and Ben Langmead. 2016. "Rail-RNA: Scalable Analysis of RNA-seq Splicing and Coverage." *Bioinformatics.* doi:10.1093/bioinformatics/btw575.

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

Karim Chine[1]
1. RosettaHUB, Inc

Keywords: Cloud Computing, Docker, Collaboration, API, Reproducible Research

Webpages: www.rosettahub.com

RosettaHUB aims at establishing a global open data science and open education meta cloud centered on usability, reproducibility, auditability, and shareability. It enables a wide range of social interactions and real-time collaborations. 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 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, 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 Pyhton 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.

Architectural Elements Enabling R for Big Data

Mark Hornick

Oracle Corporation

Keywords: Big Data, Machine Learning, Scalability, High Perforance Computing, Graph Analytics

Webpages: https://oracle.com/goto/R

Big Data garners much attention, but how can enterprises extract value from data as found in the growing corporate *data lakes* or *data reservoirs.* Extracting value from big data requires high performance and scalable tools – both in hardware and software. Increasingly, enterprises take on massive machine learning and graph analytics projects, where the goal is to build models and analyze graphs involving multi-billion row tables or to partition analyses into thousands or even millions of components.

Data scientists need to address use cases that range from modeling individual customer behavior to understanding aggregate behavior, or exploring centrality of nodes within a graph to monitoring sensors from the Internet of Things for anomalous behavior. While *R* is cited as the most used statistical language, limitations of scalability and performance often restrict its use for big data. In this talk, we present architectural elements enabling high performance and scalability, highlighting scenarios both on Hadoop/Spark and database platforms using *R*. We illustrate how **Oracle Advanced Analytics' Oracle R Enterprise** component and **Oracle R Advanced Analytics for Hadoop** enable using R on big data, achieving both scalability and performance.

useR! 2017 Oral Presentation

Improving DBI

Kirill Müller[1]

1. IVT, ETH Zurich

Keywords: Database, SQLite, specification, test suite

Webpages: https://CRAN.R-project.org/package=DBI, https://CRAN.R-project.org/package=DBItest, https://CRAN.R-project.org/package=RSQLite

Getting data in and out of R is a minor but very important part of a statistician's or data scientist's work. Sometimes, the data are packaged as R package; however, in the majority of cases one has to deal with third-party data sources. Using a database for storage and retrieval is often the only feasible option with today's ever-growing data.

DBI is R's native *DataBase Interface*: a set of virtual functions declared in the **DBI** package. DBI *backends* connect R to a particular database system by implementing the methods defined in **DBI** and accessing DBMS-specific APIs to perform the actual query processing. A common interface is helpful for both users and backend implementers. Thanks to generous support from the R Consortium, the contract for DBI's virtual functions is now specified in detail in their documentation, which are also linked to corresponding backend-agnostic tests in the **DBItest** package. This means that the compatibility of backends to the DBI specification can be verified automatically. The support from the R Consortium also allowed to bring one existing DBI backend, **RSQLite**, on par with the specification; the **odbc** package, a DBI-compliant interface to ODBC, has been written from scratch against the specification defined by **DBItest**.

Among other topics, the presentation will introduce new and updated **DBI** methods, show the design and usage of the test suite, and describe the changes in the **RSQLite** implementation.

R-based computing with big data on disk

Kylie A. Bemis[1] *and Olga Vitek*[1,2]

1. College of Computer and Information Science, Northeastern University
2. College of Science, Northeastern University

Keywords: big data, reproducibility, data aggregation, bioinformatics, imaging

Webpages: https://github.com/kuwisdelu/matter, http://bioconductor.org/packages/release/bioc/html/matter.html

A common challenge in many areas of data science is the proliferation of large and heterogeneous datasets, stored in disjoint files and specialized formats, and exceeding the available memory of a computer. It is often important to work with these data on a single machine, e.g. to quickly explore the data, or to prototype alternative analysis approaches on limited hardware. Current solutions for working with such data on disk on a single machine in *R* involve wrapping existing file formats and structures (e.g., NetCDF, HDF5, database approaches, etc.) or converting them to very simple flat files (e.g., **bigmemory**, **ff**).

Here we argue that it is important to enable more direct interactions with such data in *R*. Direct interactions avoid the time and storage cost of creating converted files. They minimize the loss of information that can occur during the conversion, and therefore improve the accuracy and the reproducibility of the analytical results. They can best leverage the rich resources from over 10,000 packages already available in *R*.

We present **matter**, a novel paradigm and a package for direct interactions with complex, larger-than-memory data on disk in *R*. **matter** provides transparent access to datasets on disk, and allows us to build a single dataset from many smaller data fragments in custom formats, without reading them into memory. This is accomplished by means of a flexible data representation that allows the structure of the data in memory to be different from its structure on disk. For example, what **matter** presents as a single, contiguous vector in *R* may be composed of many smaller fragments from multiple files on disk. This allows **matter** to scale to large datasets, stored in large stand-alone files or in large collections of smaller files.

To illustrate the utility of **matter**, we will first compare its performance to **bigmemory** and **ff** using data in flat files, which can be easily accessed by all the three approaches. In tests on simulated datasets greater than 1 GB and common analyses such as linear regression and principal components analysis, **matter** consumed the same or less memory, and completed the analyses in a comparable time. It was therefore similar or more efficient than the available solutions.

Next, we will illustrate the advantage of **matter** in a research area that works with complex formats. Mass spectrometry imaging (MSI) relies on imzML, a common open-source format for data representation and sharing across mass spectrometric vendors and workflows. Results of a single MSI experiment are typically stored in multiple files. An integration of **matter** with the *R* package **Cardinal** allowed us to perform statistical analyses of all the datasets in a public Gigascience repository of MSI datasets, ranging from <1 GB up to 42 GB in size. All of the analyses were performed on a single laptop computer. Due to the structure of imzML, these analyses would not have been possible with the existing alternative solutions for working with larger-than-memory datasets in *R* .

Finally, we will demonstrate the applications of **matter** to large datasets in other formats, in particular text data that arise in applications in genomics and natural language processing, and will discuss approaches to using **matter** when developing new statistical methods for such datasets.

Programming with tidyverse grammars

Lionel Henry A[1] and Hadley Wickham B[1]

1. RStudio

Keywords: tidyeval, tidyverse, dplyr, quasiquotation, NSE

Webpages: https://CRAN.R-project.org/package=dplyr, https://github.com/hadley/rlang

Evaluating code in the context of a dataset is one of R's most useful feature. This idiom is used in base R functions like `subset()` and `transform()` and has been developed in tidyverse packages like dplyr and ggplot2 to design elegant grammars. The downside is that such interfaces are notoriously difficult to program with. It is not as easy as it should be to program with dplyr inside functions in order to reduce duplicated code involving dplyr pipelines. To solve these issues, RStudio has developed *tidyeval*, a set of new language features that make it straightforward to program with these grammars. We present tidyeval in this talk with a focus on solving concrete problems with popular tidyverse packages like dplyr.

Distributional Trees and Forests

Lisa Schlosser[1], *Torsten Hothorn*[2], *Achim Zeileis*[1]

1. *University of Innsbruck*
2. *University of Zurich*

Keywords: Distributional regression, recursive partitioning, decision trees, random forests

Webpages: https://R-Forge.R-project.org/projects/partykit/

In regression analysis one is interested in the relationship between a dependent variable and one or more explanatory variables. Various methods to fit statistical models to the data set have been developed, starting from ordinary linear models considering only the mean of the response variable and ranging to probabilistic models where all parameters of a distribution are fit to the given data set.

If there is a strong variation within the data it might be advantageous to split the data first into more homogeneous subgroups based on given covariates and then fit a local model in each subgroup rather than fitting one global model to the whole data set. This can be done by applying regression trees and forests.

Both of these two concepts, parametric modeling and algorithmic trees, have been investigated and developed further, however, mostly separated from each other. Therefore, our goal is to embed the progress made in the field of probabilistic modeling in the idea of algorithmic tree and forest models. In particular, more flexible models such as GAMLSS (Rigby and Stasinopoulos 2005) should be fitted in the nodes of a tree in order to capture location, scale, shape as well as censoring, tail behavior etc. while non-additive effects of the explanatory variables can be detected by the splitting algorithm used to build the tree.

The corresponding implementation is provided in an *R* package **disttree** which is available on R-Forge and includes the two main functions `disttree` and `distforest`. Next to the data set and a formula the user only has to specify a distribution family and receives a tree/forest model with a set of distribution parameters for each final node. One possible way to specify a distribution family is to hand over a **gamlss.dist** family object (Stasinopoulos, Rigby, and others 2007). In `disttree` and `distforest` the fitting function `distfit` is applied within a tree building algorithm chosen by the user. Either the MOB algorithm, an algorithm for model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008), or the ctree algorithm (Hothorn, Hornik, and Zeileis 2006) can be used as a framework. These algorithms are both implemented in the **partykit** package (Hothorn et al. 2015).

References

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3). Taylor & Francis: 651–74.

Hothorn, Torsten, Kurt Hornik, Carolin Strobl, and Achim Zeileis. 2015. "Package 'Party'." *Package Reference Manual for Party Version 0.9–0.998* 16: 37.

Rigby, Robert A, and D Mikis Stasinopoulos. 2005. "Generalized Additive Models for Location Scale and Shape (with Discussion)." *Applied Statistics* 54.3: 507–54.

Stasinopoulos, D Mikis, Robert A Rigby, and others. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software* 23 (7): 1–46.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17 (2). Taylor & Francis: 492–514.

Community-based learning and knowledge sharing

Lahti L[1], Lehtomäki J[2], and Kainu M^3
1. Department of Mathematics and Statistics, University of Turku, Finland
2. Department of Earth Sciences, VU Amsterdam
3. The Social Insurance Institution of Finland

Keywords: Teaching, Knowledge Sharing, Best Practices

Webpage: https://github.com/rOpenGov/edu

R is increasingly used to teach programming, quantitative analytics, and reproducible research practices. Based on our combined experience from universities, research institutes, and the public sector, we summarize key ingredients for teaching of modern data science. Learning to program has already been greatly facilitated by initiatives such as Data Carpentry and Software Carpentry, and educational resources have been developed by the users, including domain specific tutorial collections and training materials (Kamvar et al. 2017; Lahti et al. 2017; Afanador-Llach et al. 2017). An essential pedagogical feature of R is that it enables a problem-centered, interactive learning approach. Even programming-naive learners can, in our experience, rapidly adopt practical skills by analyzing topical example data sets supported by ready-made Rmarkdown templates; these can provide an immediate starting point to expose the learners to some of the key tools and best practices (Wilson et al. 2016). However, many aspects of learning R are still better appreciated by advanced users; such as harnessing the full potential of open collaboration model by joint development of custom R packages, report templates, shiny-modules, or database functions that enables rapid development of solutions catering specific practical needs. Indeed, at all levels of learning, getting things done fast, appears to be an essential component for successful learning as it provides instant rewards and helps to put the acquired skills into immediate use. The diverse needs of different application domains pose a great challenge for crafting common guidelines and materials, however. Leveraging the existing experience within the learning community can greatly support the learning process as it helps to ensure the domain specificity and relevance of the teaching experience. This can actively promoted by peer support and knowledge sharing; some ways to achieve this include code review, show-and-tell culture, informal meetings, online channels (e.g. Slack, IRC, Facebook) and hackathons. Last but not least, having fun throughout the learning process is essential; gamification of assignments with real-time rankings or custom functions performing non-statistical operations like emailing gif images can raise awareness of how R as a full-fledged programming language differs from proprietary statistical packages. In order to meet these demands, we designed specific open infrastructure to support learning in R. Our infrastructure gathers a set of modules to construct domain spesific assignments for various phases of data analysis. The assignments are coupled with automated evaluation and scoring routines that provide instant feedback during learning. In this talk, we introduce these R-based teaching tools and summarize our practical experiences on the various pedagogical aspects, opportunities, and challenges of community-based learning and knowledge sharing enabled by the R ecosystem.

References

Afanador-Llach, Maria José, Antonio Rojas Castro, Adam Crymble, Víctor Gayol, Fred Gibbs, Caleb McDaniel, Ian Milligan, Amanda Visconti, and Jeri Wieringa. 2017. "The Programming Historian. Second Edition." http://programminghistorian.org/.

Kamvar, Zhian N., Margarita M. López-Uribe, Simone Coughlan, Niklaus J. Grünwald, Hilmar Lapp, and Stéphanie Manel. 2017. "Developing Educational Resources for Population Genetics in R: An Open and Collaborative Approach." *Molecular Ecology Resources* 17 (1): 120–28. doi:10.1111/1755-0998.12558.

Lahti, Leo, Sudarshan Shetty, Tineka Blake, and Jarkko Salojarvi. 2017. "Microbiome R Package." http://microbiome.github.io/microbiome.

Wilson, Greg, Jenny Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy Teal. 2016. "Good Enough Practices for Scientific Computing," 1–30.

papr: Tinder for pre-prints, a Shiny Application for collecting gut-reactions to pre-prints from the scientific community

Lucy D'Agostino McGowan[1], Nick Strayer[1], and Jeff Leek[2]

1. *Vanderbilt University*
2. *Johns Hopkins University*

Keywords: shiny, recommender engine, social network, shinysense

Webpages: https://jhubiostatistics.shinyapps.io/papr/, https://github.com/lfod/papr

papr is an R Shiny web application and social network for evaluating bioRxiv pre-prints. The app serves multiple purposes, allowing the user to quickly swipe through pertinent abstracts as well as find a community of researchers with similar interests. It also serves as a portal for accessible "open science", getting abstracts into the hands of users of all skill levels. Additionally, the data could help build a general understanding of what research the community finds exciting.

We allow the user to log in via Google to track multiple sessions and have implemented a recommender engine, allowing us to tailor which abstracts are shown based on each user's previous abstract rankings. While using the app, users view an abstract pulled from bioRxiv and rate it as "exciting and correct", "exciting and questionable", "boring and correct", or "boring and questionable" by swiping the abstract in a given direction. The app includes optional social network features, connecting users who provide their twitter handle to users who enjoy similar papers.

This presentation will demonstrate how to incorporate tactile interfaces, such as swiping, into a Shiny application using a package we created for this functionality shinysense, store real-time user data on Dropbox using drop2, login in capabilities using googleAuthR and googleID, how to implement a recommender engine using principle component analysis, and how we have handled issues of data safety/security through proactive planning and risk mitigation. Finally, we will report the app activity, summarizing both the user traffic and what research users are finding exciting.

addhaz: Contribution of chronic diseases to the disability burden in *R*

Renata T. C. Yokota [1,2], *Caspar W. N. Looman* [3], *Wilma J. Nusselder* [3], *Herman Van Oyen* [1,4], *Geert Molenberghs* [5,6]

1. *Department of Public Health and Surveillance, Scientific Institute of Public Health, Brussels, Belgium*
2. *Department of Sociology, Interface Demography, Vrije Universiteit Brussel, Brussels, Belgium*
3. *Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*
4. *Department of Public Health, Ghent University, Ghent, Belgium*
5. *Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Universiteit Hasselt, Diepenbeek, Belgium*
6. *Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium*

The increase in life expectancy followed by the growing proportion of old individuals living with chronic diseases contributes to the burden of disability worldwide. The estimation of how much each chronic condition contributes to the disability prevalence can be useful to develop public health strategies to reduce the burden. In this presentation, we will introduce the *R* package **addhaz**, which is based on the attribution method (Nusselder and Looman 2004) to partition the total disability prevalence into the additive contributions of chronic diseases using cross-sectional data. The *R* package includes tools to fit the binomial and multinomial additive hazard models, the core of the attribution method. The models are fitted by maximizing the binomial and multinomial log-likelihood functions using constrained optimization (`constrOptim`). The 95% Wald and bootstrap percentile confidence intervals can be obtained for the parameter estimates. Also, the absolute and relative contribution of each chronic condition to the disability prevalence and their bootstrap confidence intervals can be estimated. An additional feature of **addhaz** is the possibility to use parallel computing to obtain the bootstrap confidence intervals, reducing computation time. In this presentation, we will illustrate the use of **addhaz** with examples for the binomial and multinomial models, using the data from the Brazilian National Health Survey, 2013.

Keywords: Disability, Binomial outcome, Multinomial outcome, Additive hazard model, Cross-sectional data

Webpage: https://cran.r-project.org/web/packages/addhaz/index.html

References

Nusselder, Wilma J, and Caspar WN Looman. 2004. "Decomposition of Differences in Health Expectancy by Cause." *Demography* 41 (2). Springer: 315–34.

Stochastic Gradient Descent Log-Likelihood Estimation in the Cox Proportional Hazards Model with Applications to The Cancer Genome Atlas Data

Marcin Kosinski[1]

1. Grupa Wirtualna Polska.

Keywords: Log-Likelihood Estimation, Stochastic Gradient Descent, Cox PH Model, Survival Analysis

Webpages: https://CRAN.R-project.org/package=coxphSGD, https://github.com/MarcinKosinski/coxphSGD

In the last decade, the volume of data have grown faster then the speed of processors. In this situation the statistical machine learnig methods have become more limited by the computations time than the volume of datasets. Compromise solutions in the case of large scale data are associated with the computational complexity of optimization methods, which must be made in a non-trivial way. One of such solutions are optimization algorithms that are basen on a stochastic gradient descent (Bottou (2010), Bottou (2012), Widrow (1960)), which exhibit a high efficiency during operations on the data of a large scale.

In my presentation I will describe the stochastic gradient descent algorithm that was applied in the log-likelihood estimation process of coefficients' calcualtions of the Cox proportional hazards model. This algorithm can be successfully used in a time to event analyzes, in which which the number of explanatory variables significantly exceeds the number of observations. The prepared method of estimation of coefficients with the usage of a stochastic gradient decent can be applied in survival analyzes from ares like: molecular biology, bioinformatical screenings of gene expressions or analyzes based on DNA microarrays, that are widely used in the clinical diagnostics, treatment and research.

The created estimation workflow was a new approach (in the time I wrote my master thesis), not known in the literature. It's resistant to the problem of variables collinearity and works well in situations of continuous coefficients improvement for a streaming data.

References

Bottou, L. 2010. "Large-Scale Machine Learning with Stochastic Gradient Descent."

———. 2012. "Stochastic Gradient Descent Tricks."

Widrow, B. 1960. An Adaptive ADALINE Neuron Using Chemical Memistors. Technical Report No. 1553-2, Stanford University.

Statistics in Action with R: an educative platform

Marc Lavielle[1,2]

1. Inria Saclay Ile-de-France
2. Ecole Polytechnique

Keywords: Hypothesis testing, regression model, mixed effects model, mixture model, change point detection

Webpage: http://sia.webpopix.org/

We are developing at Inria and Ecole Polytechnique the web-based educative platform *Statistics in Action with R*.
The purpose of this online course is to show how statistics may be efficiently used in practice using *R*.

The course presents both statistical theory and practical analysis on real data sets. The *R* statistical software and several *R* packages are used for implementing methods presented in the course and analyzing real data. Many interactive **Shiny apps** are also available.

Topics covered in the current version of the course are:

- hypopthesis testing (single and multiple comparisons)
- regression models (linear and nonlinear models)
- mixed effects models (linear and nonlinear models)
- mixture models
- detection of change points
- image restoration

We are aware that important aspects of statistics are not addressed, both in terms of models and methods. We plan to fill some of these gaps shortly.

Even if *R* is extensively used for this course, this is not a *R* programming course. On one hand, our objective is not to propose the most efficient implementation of an algorithm, but rather to provide a code that is easy to understand, to reuse and to extend.

On the other hand, the *R* functions used to illustrate a method are not used as "black boxes". We show in detail how the results of a given function are obtained. Then, the course may be read at two different levels: we may be only interested in the statistical technique to use (and then the *R* function to use) for a given problem (see the first part of the course about polynomial regression), or we may want to go into details and understand how these results are computed (see the second part of this course about polynomial regression).

This course was first given at Ecole Polytechnique (France) in 2017.

useR! 2017 Oral Presentation

shiny.collections: Google Docs-like live collaboration in Shiny

Marek Rogala[1] and Filip Stachura[1]

1. Appsilon Data Science

Keywords: Shiny, data applications, UX, live collaboration, data persistence

Webpages: https://appsilon.github.io/shiny.collections/

What users expect from web applications today differs dramatically from what was available 5 years ago. They are used to interactivity, data persistence, and what's more, the ability to share live collaboration experiences, like in *Google Docs*. If one user changes the data, other users want to see the changes immediately on their screens. They don't care whether it is a data-exploration app from a data scientist or a solution built by a team of software engineers.

Shiny is perfect for building interactive data-driven applications suited for the modern user. In this presentation, we show how to create real-time collaboration experience in **Shiny** apps.

From the presentation, you will learn the concepts of reactive databases, how to use them in **Shiny**, and how to adapt existing components to provide live collaboration.

We will present a package we developed for that. **shiny.collections** adds persistent reactive collections that can be effortlessly integrated with components like **Shiny** inputs, **DT** data table or **rhandsontable**. The package makes it easy to build collaborative **Shiny** applications with persistent data.

The presentation will be very actionable. Our goal is for everyone in the audience to be able to add persistence and collaboration to their apps in less than 10 minutes.

Integrated Cluster Analysis with R in Drug Discovery Experiments using Multi-Source Data

Marijke Van Moerbeke[1], *Nolen-Joy Perualila-Tan*[1], *Ziv Shkedy*[1] *and Dhammika Amaratunga*[2]

1. *Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Belgium*
2. *Independent consultant*

Keywords: High dimensional data, Clustering

Webpages: https://cran.r-project.org/web/packages/IntClust/index.html

Discovering the exact activities of a compound is of primary interest in drug development. A single drug can interact with multiple targets and unintended drug-target interactions could lead to severe side effects. Therefore, it is valuable in the early phases of drug discovery to not only demonstrate the desired on-target efficacy of compounds but also to outline its unwanted off-target effects. Further, the earlier unwanted behaviour is documented, the better. Otherwise, the drug could fail in a later stage which means that the invested time, effort and money are lost.

In the early stages of drug development, different types of information on the compounds are collected: the chemical structures of the molecules (fingerprints), the predicted targets (target predictions), on various bioassays, the toxicity and more. An analysis of each data source could reveal interesting yet disjoint information. It only provides a limited point of view and does not give information on how everything is interconnected in the global picture (Shi, De Moor, and Moreau 2009). Therefore, a simultaneous analysis of multiple data sources can provide a more complete insight on the compounds' activity.

An analysis based on multiple data sources is relatively new and growing area in drug discovery and drug development. Multi-source clustering procedures provide us with the opportunity to relate several data sources to each other to gain a better understanding of the mechanism of action of compounds. The use of multiple data sources was investigated in the QSTAR (quantitative structure transcriptional activity relationship) consortium (Ravindranath et al. 2015). The goal was to find associations between chemical, bioassay and transcriptomic data in the analysis of a set of compounds under development.

In the current study, we extend the clustering method presented in(Perualila-Tan et al. 2016) and review the performance of several clustering methods on a real drug discovery project in R. We illustrate how the new clustering approaches provide a valuable insight for the integration of chemical, bioassay and transcriptomic data in the analysis of a specific set of compounds. The proposed methods are implemented and publicly available in the R package IntClust which is a wrapper package for a multitude of ensemble clustering methods.

References

Perualila-Tan, N., Z. Shkedy, W. Talloen, H. W. H. Goehlmann, QSTAR Consortium, M. Van Moerbeke, and A. Kasim. 2016. "Weighted-Similarity Based Clustering of Chemical Structure and Bioactivity Data in Early Drug Discobased." *Journal of Bioinformatics and Computational Biology.*

Ravindranath, A. C., N. Perualila-Tan, A. Kasim, G. Drakakis, S. Liggi, S. C. Brewerton, D. Mason, et al. 2015. "Connecting Gene Expression Data from Connectivity Map and in Silico Target Predictions for Small Molecule Mechanism-of-Action Analysis." *Mol. BioSyst.* 11 (1). The Royal Society of Chemistry: 86–96. doi:10.1039/C4MB00328D.

Shi, Y., B. De Moor, and Y. Moreau. 2009. "Clustering by Heterogeneous Data Fusion: Framework and Applications." *NIPS Workshop.*

Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data with ctmcd

Marius Pfeuffer

Department of Statistics and Econometrics, University of Erlangen-Nuremberg

Keywords: Embedding Problem, Generator Matrix, Continuous-Time Markov Chain, Discrete-Time Markov Chain

Webpages: https://CRAN.R-project.org/package=ctmcd

The estimation of the parameters of a continuous-time Markov chain from discrete-time data is an important statistical problem which occurs in a wide range of applications: e.g., with the analysis of gene sequence data, for causal inference in epidemiology, for describing the dynamics of open quantum systems in physics, or in rating based credit risk modeling to name only a few.

The parameters of a continuous-time Markov chain are called generator matrix (also: transition rate matrix or intensity matrix) and the issue of estimating generator matrices from discrete-time data is also known as the embedding problem for Markov chains. For dealing with this missing data situtation, a variety of estimation approaches have been developed. These comprise adjustments of matrix logarithm based candidate solutions of the aggregated discrete-time data, see (Israel, Rosenthal, and Wei 2001) or (Kreinin and Sidelnikova 2001). Moreover, likelihood inference can be conducted by an instance of the expectation-maximization (EM) algorithm and Bayesian inference by a Gibbs sampling procedure based on the conjugate gamma prior distribution (Bladt and Sørensen 2005).

The *R* package **ctmcd** (Pfeuffer 2016) is the first publicly available implementation of the approaches listed above. Besides point estimates of generator matrices, the package also contains methods to derive confidence and credibility intervals. The capabilities of the package are illustrated using Standard & Poor's discrete-time credit rating transition data. Moreover, methodological issues of the described approaches are discussed, i.e., the derivation of the conditional expectations of the E-Step in the EM algorithm and the sampling of endpoint-conditioned continuous-time Markov chain trajectories for the Gibbs sampler.

References

Bladt, M., and M. Sørensen. 2005. "Statistical Inference for Discretely Observed Markov Jump Processes." *Journal of the Royal Statistical Society B.*

Israel, R. B., J. S. Rosenthal, and J. Z. Wei. 2001. "Finding Generators for Markov Chains via Empirical Transition Matrices, with Applications to Credit Ratings." *Mathematical Finance.*

Kreinin, A., and M. Sidelnikova. 2001. "Regularization Algorithms for Transition Matrices." *Algo Research Quarterly.*

Pfeuffer, M. 2016. "ctmcd: An R Package for Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data." *In Revision (the R Journal).*

useR! 2017 Oral Presentation

RL10N: Translating Error Messages & Warnings

Richie Cotton[1] and Thomas Leeper[2]

1. DataCamp
2. Department of Government, London School of Economics

Keywords: package development, localization, translation, errors and warnings, R Consortium

Webpages: https://CRAN.R-project.org/package=msgtools, https://github.com/RL10N

R is becoming the global standard language for data analysis, but it requires its user to speak English. RL10N is an R Consortium funded project to make it easier to translate error messages and warnings into different languages. The talk covers how to automatically translate messages using Google Translate and Microsoft Translator, and how to easily integrate these message translations into your R packages using `msgtools`. Make your code more accessible to users around the world!

Maps are data, so why plot data on a map?

Mark Padgham[1]

1. Department of Geoinformatics, University of Salzburg, Austria

Keywords: data maps, OpenStreetMap, spatial, visualization

Webpages: https://CRAN.R-project.org/package=osmplotr, https://github.com/ropensci/osmplotr, https://github.com/osmdatar/osmdata

R, like any and every other system for analysing and visualising spatial data, has a host of ways to overlay data on maps (or the other way around). Maps nevertheless contain data—nay, maps *are* data—making this act tantamount to overlaying data upon data. That's likely not going to end well, and so this talk will present two new packages that enable you to visualise your own data with actual map data such as building polygons or street lines, rather than merely overlaying (or underlaying) them. The **osmdata** package enables publically accessible data from OpenStreetMap to be read into *R*, and **osmplotr** can then use these data as a visual basis for your own data. Both categorical and continuous data can be visualised through colours or through structural properties such as line thicknesses or types. We think this results is more visually striking and beautiful data maps than any alternative approach that necessitates separating your data from map data.

Morphological Analysis with R

Markus Voge[1]

1. EA European Academy of Technology and Innovation Assessment

Keywords: Shiny, DataTables, analysis methods, problem-solving

Webpages: https://github.com/sgrubsmyon/morphr

Morphological analysis is a problem-structuring method developed by the astrophysicist Fritz Zwicky in the 1940s to 1960s [1–3]. It can be used to explore and constrain a multi-dimensional, possibly non-quantifiable, problem space. The problem is put into a *morphological field*, a tabular representation where each parameter corresponds to a column whose rows are filled with the parameter values. Each parameter value is mutually checked for consistency with all other parameter values. This enables to systematically exclude inconsistent configurations and therefore greatly reduces the problem space.

Dedicated software is helpful to visualize and work with a morphological field. While full-fledged software solutions already exist [4,5], they are confined to the Windows desktop and cannot be run in a web browser. The R package **morphr** is a first step into the direction of a browser-based morphological analysis tool. By leveraging *R* technology, one can relatively easily bring morphological analysis into a modern, web-centric, cross-platform environment, embedded in an open source ecosystem. Morphological fields and their constraints can be visualized interactively in a web browser: a user can select parameter values via mouse click, causing the field to highlight the remaining configurations consistent with the selection. To provide the interactivity, the package is using *R*'s **shiny** package and is built on top of the excellent **DT** package, which is an *R* wrapper around the *JavaScript* library **DataTables**.

In this talk, morphological analysis in general is introduced and it is shown how a morphological analysis can be assisted with *R* using **morphr**.

References

1. Zwicky F. Morphology and nomenclature of jet engines. *Aeronautical Engineering Review* (1947) **6**:49–50.

2. Zwicky F. Morphological astronomy. *The Observatory* (1948) **68**:121–143. Available at: http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1948Obs....68..121Z&amp;data_type=PDF_HIGH&amp;whole_paper=YES&amp;type=PRINTER&amp;filetype=.pdf

3. Zwicky F, Wilson AG. *New methods of thought and procedure - contributions to the symposium on methodologies*. Berlin Heidelberg: Springer Science & Business Media (1967). doi:10.1007/978-3-642-87617-2

4. Swedish Morphological Society. General Morphological Analysis - A general method for non-quantified modeling. (2002 (Revised 2013)) Available at: http://www.swemorph.com/ma.html

5. Swedish Morphological Society. MA/Carma™ - Advanced Computer Support for General Morphological Analysis. (2005–2016) Available at: http://www.swemorph.com/macarma.html

brms: Bayesian Multilevel Models using Stan

Paul Bürkner[1]

1. Institute of Psychology, University of Münster, Germany

Keywords: multilevel models, Bayesian inference, Stan

Webpages: https://cran.r-project.org/web/packages/brms, https://github.com/paul-buerkner/brms

The **brms** package (Bürkner, in press) implements Bayesian multilevel models in R using the probabilistic programming language *Stan* (Carpenter, 2017). A wide range of distributions and link functions are supported, allowing users to fit linear, robust linear, binomial, Poisson, survival, response times, ordinal, quantile, zero-inflated, hurdle, and even non-linear models all in a multilevel context. Further modeling options include auto-correlation and smoothing terms, user defined dependence structures, censored data, meta-analytic standard errors, and quite a few more. In addition, all parameters of the response distribution can be predicted in order to perform distributional regression. Prior specifications are flexible and explicitly encourage users to apply prior distributions that actually reflect their beliefs. In addition, model fit can easily be assessed and compared with posterior predictive checks and leave-one-out cross-validation.

References

Bürkner, P. C. (in press). **brms**: An R Package for Bayesian Multilevel Models using Stan. *Journal of Statistical Software.*

Carpenter B., et al. (2017). *Stan:* A Probabilistic Programming Language. *Journal of Statistical Software.*

useR! 2017 Oral Presentation

Interactive bullwhip effect exploration using SCperf and Shiny

Marlene Silva-Marchena

Independent consultant

Abstract: The bullwhip effect, an increase in demand variability along the supply chain, is pointed out as a key driver of inefficiencies associated with the supply chain. In the presence of this phenomenon, participants involved in the manufacture of a product and its distribution to final customer face unstable production schedules or excessive inventory.

Although there are several implementations illustrating the bullwhip effect, it still remains difficult for scholars and supply chain practitioners to understand and quantify its real effect in the suply chain performance.

Using the **SCperf** package and **Shiny** app, we have developed an interactive bullwhip game which follows the standard setup of the classic "Beer Distribution Game" (MIT). Our web interface illustrates the distribution process of a multi-echelon supply chain, the goal of the game being to minimize costs along the chain while satisfying service level requirements.

Our open source application is user friendly, easily supports sophisticated forecasting techniques and inventory models and does not require any *R* experience.

In this talk, we describe the underlying design of the game and show by means of examples how changing the forecasting method or tuning the parameters of the replenishment policy (lead time, customer service level, etc) induces or reduces the bullwhip effect. This application may be adapted to become a learning tool in classroom and training programs.

Keywords: bullwhip effect, beergame, inventory, SCperf, shiny

Ensemble packages with user friendly interface: an added value for the *R* community

Martin Otava[1], Rudradev Sengupta[2], Ewoud de Troyer[2] and Ziv Shkedy[2]

1. Manufacturing, Toxicology and Applied Statistical Sciences, Janssen Research & Development, Janssen Pharmaceutica NV, Turnhoutseweg 30, B-2340 Beerse, Belgium
2. Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Martelarenlaan 42, B-3500 Hasselt, Belgium

Keywords: Ensemble package, user-friendly interface, gene expression analysis, biclustering

Webpages: https://r-forge.r-project.org/R/?group_id=589, https://github.com/ewouddt/RcmdrPlugin.BiclustGUI

The increasing amount of *R* packages makes it difficult to any newcomer to orientate himself/herself in the large amount of option available for topics such as modeling, clustering, variable selection, optimization, sample size estimation etc. The quality of the packages, the associated help files, error reporting system and continuity of support vary significantly and methods may be duplicated across multiple packages if the packages focus on a specific application within a particular field only.

Ensemble packages can be seen as another type of contribution to the *R* community. Careful revision of packages that approach the same topic from different perspectives may be very useful for increasing the overall quality of the *CRAN* repository. The revision should not be limited to the technical part, but should also cover methodological aspects. A necessary condition for success of the ensemble package is of course that this revision happens in close collaboration with the authors of the original package.

An additional benefit of ensemble packages lies in leveraging many graphical options of the traditional *R* framework. Starting from a simple Graphical User Interface, over an *R Commander* plugins, to *Shiny* applications, *R* provides wide range of visualization options. By combining visualization with the content of original packages, the ensemble package can provide different user experience. Such a property extends added value of ensemble beyond a simple review library. Necessarily, the flexibility of the package is reduced by transformation into point and click interface, but the user requiring a fully flexible environment can be referred to the original packages.

We present two case studies of such ensemble packages: **IsoGeneGUI** and **BiclustGUI**. **IsoGeneGUI** is implemented in the Graphical User Interface (*GUI*) and combines the original **IsoGene** package for dose-response analysis of high dimensional data with other packages such as **orQA**, **ORIClust**, **goric** and **ORCME**, that offer methods to analyze different perspectives of gene expression based data sets. **IsoGeneGUI** thus provides a wide range of methods methods (and the most complete data analysis tool for order restricted analysis) in a user friendly fashion. Hence analyzes can be implemented by users with only limited knowledge of *R* programming. The **RcmdrPlugin.BiclustGUI** is a *GUI* plugin for *R Commander* that combines various biclustering packages, bringing multiple algorithms, visualizations and diagnostics tools into one unified framework. Additionally, the package allows for simple inclusion of potential future biclustering methods.

The collaboration with the authors of the original packages on implementation of their methods within an ensemble package was extremely important for both case studies. Indeed, in that way, the link with the original packages could be retained. The ensemble package allowed for careful evaluation of the methods, their overlap and differences, and for presenting them as a concise framework in a user friendly environment.

useR! 2017 Oral Presentation

Biosignature-Based Drug Design: from high dimensional data to business impact

Marvin Steijaert[1], Griet Laenen[1], Joerg Kurt Wegner[2], Vladimir Chupakhin[2], José Felipe Golib Dzib[2] and Hugo Ceulemans[2]

1. Open Analytics NV
2. Janssen Pharmaceutica

Keywords: biosignatures, machine learning, drug design, data fusion, high-throughput screening

Webpages: https://www.openanalytics.eu/

For decades, high throughput screening of chemical compounds has played a central role in drug design. In general, such screens were only affordable if they had a narrow biological scope (e.g., compound activity on an isolated protein target). In recent years, screening techniques have become available that combine a high throughput with a high dimensional readout and a complex biological context (e.g., cell culture). Examples are high content imaging and L1000 transcriptomics. In addition, due to state-of-the-art machine learning methods (Unterthiner et al. 2014) and high performance computing (Harnie et al. 2016) it has become possible to benefit from such high dimensional biological data on an enterprise scale. Together, these advances enable *Biosignature-Based Drug Design*, a paradigm that will dramatically change pharmaceutical research.

A software pipeline, mainly built in *R* and *C++*, allows us to support Biosignature-Based Drug Design in an enterprise setting. It is worth noting that dealing with multiple data sets of this scale and complexity is non-trivial and challenging. With our pipeline, we tailor generic methods to the needs of specific projects in diverse therapeutic areas. This operational application goes hand in hand with an ongoing effort –together with academic partners– to improve and extend our workflow.

We will show use cases in which Biosignature-Based Drug Design has increased the effectiveness and cost-efficiency of high throughput screens by repurposing historic data (Simm et al. 2017). Moreover, integrating multiple data sources allows to takes into account a broader biological context, rather than a single mode of action. This will yield a better understanding of on- and off-target effects. Ultimately, this may reduce failure rates for drug candidates in clinical trials.

Acknowledgements

This work was supported by research grants IWT130405 ExaScience Life Pharma and IWT150865 Exaptation from the Flanders Innovation and Entrepreneurship agency.

References

Harnie, D., M. Saey, A. E. Vapirev, J.K. Wegner, A. Gedich, M.N. Steijaert, H. Ceulemans, R. Wuyts, and W. De Meuter. 2016. "Scaling Machine Learning for Target Prediction in Drug Discovery Using Apache Spark." *Future Generation Computer Systems.*

Simm, J., G. Klambauer, A. Arany, M.N. Steijaert, J.K. Wegner, E. Gustin, V. Chupakhin, et al. 2017. "Repurposed High-Throughput Images Enable Biological Activity Prediction for Drug Discovery." *bioRxiv.*

Unterthiner, T., A. Mayr, G. Klambauer, M.N. Steijaert, H. Ceulemans, J.K. Wegner, and S. Hochreiter. 2014. "Deep Learning as an Opportunity in Virtual Screening." In *Workshop on Deep Learning and Representation Learning (Nips 2014).*

A quasi-experiment for the influence of the user interface on the acceptance of R

Matthias Gehrke[1] *and Karsten Luebke*[1]

1. FOM University of Applied Sciences

Keywords: Teaching Statistics using R, User Interface, Technology Acceptance Model

Teaching computation with *R* within statistic education is affected by the acceptance of the technology (Baglin 2013). In a quasi-experiment we investigate whether different user interfaces to *R*, namely **mosaic** or **Rcmdr**, influence the acceptance according to the Technology Acceptance Model (Venkatesh et al. 2003).

The focus thereby is on the perceived usefulness and ease of use of *R* software for people studying while working in different economy related disciplines. At our private university of applied science for professionals studying while working with more than 30 study centres across Germany use of *R* is compulsory in all statistical courses in all the different Master programs and in all study centres. Due to a change in the course of study we were able to teach the lecture twice in one term, one with **Rcmdr**, one with **mosaic**, enabling a quasi-experimental setup for two lectures each.

References

Baglin, James. 2013. "Applying a Theoretical Model for Explaining the Development of Technological Skills in Statistics Education." *Technology Innovations in Statistics Education* 7 (2). https://escholarship.org/uc/item/8w97p75s.

Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. "User Acceptance of Information Technology: Toward a Unified View." *MIS Q.* 27 (3): 425–78. http://dl.acm.org/citation.cfm?id=2017197.2017202.

Actuarial and statistical aspects of reinsurance in *R*

Tom Reynkens[1], *Jan Beirlant*[1,2] *and Roel Verbelen*[3]

1. *LStat and LRisk, Department of Mathematics, KU Leuven, Belgium*
2. *Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa*
3. *LStat and LRisk, Faculty of Economics and Business, KU Leuven, Belgium*

Keywords: extreme value theory, censoring, splicing, risk measures

Webpages: https://CRAN.R-project.org/package=ReIns, https://github.com/TReynkens/ReIns

Reinsurance is an insurance purchased by one party (usually an insurance company) to indemnify parts of its underwritten insurance risk. The company providing this protection is then the reinsurer. A typical example of a reinsurance is an excess-loss insurance where the reinsurer indemnifies all losses above a certain threshold that are incurred by the insurer. Albrecher, Beirlant, and Teugels (2017) give an overview of reinsurance forms, and its actuarial and statistical aspects: models for claim sizes, models for claim counts, aggregate loss calculations, pricing and risk measures, and choice of reinsurance. The **ReIns** package, which complements this book, contains estimators and plots that are used to model claim sizes. As reinsurance typically concerns large losses, extreme value theory (EVT) is crucial to model the claim sizes. **ReIns** provides implementations of classical EVT plots and estimators (see e.g. Beirlant et al. 2004) which are essential tools when modelling heavy-tailed data such as insurance losses.

Insurance claims can take long before being completely settled, i.e. there is a long time between the occurrence of the claim and the final payment. If the claim is notified to the (re)reinsurer but not completely settled before the evaluation time, not all information on the final claim amount is available, and hence censoring is present. Several EVT methods for censored data are included in **ReIns**.

A global fit for the distribution of losses is e.g. needed in reinsurance. Modelling the whole range of the losses using a standard distribution is usually very hard and often impossible. A possible solution is to combine two distributions in a splicing model: a light-tailed distribution for the body, i.e. light and moderate losses, and a heavy-tailed distribution for the tail to capture large losses. Reynkens et al. (2016) propose a splicing model with a mixed Erlang (ME) distribution for the body and a Pareto distribution for the tail. This combines the flexibility of the ME distribution with the ability of the Pareto distribution to model extreme values. **ReIns** contains the implementation of the expectation maximisation (EM) algorithm to fit the splicing model to censored data. Risk measures and excess-loss insurance premiums can be computed using the fitted splicing model.

In this talk, we apply the plots and estimators, available in **ReIns**, to model real life insurance data. Focus will be on the splicing modelling framework and other methods adapted for censored data.

References

Albrecher, Hansjörg, Jan Beirlant, and Jef Teugels. 2017. *Reinsurance: Actuarial and Statistical Aspects*. Wiley, Chichester.

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jef Teugels. 2004. *Statistics of Extremes: Theory and Applications*. Wiley, Chichester.

Reynkens, Tom, Roel Verbelen, Jan Beirlant, and Katrien Antonio. 2016. "Modelling Censored Losses Using Splicing: A Global Fit Strategy with Mixed Erlang and Extreme Value Distributions." https://arxiv.org/abs/1608.01566.

Visual funnel plot inference for meta-analysis

Author Michael Kossmeier[1], Ulrich S. Tran[1] and Martin Voracek[1]

1. Department of Basic Psychological Research and Research Methods, School of Psychology, University of Vienna, Austria

Keywords: meta-analysis, funnel plot, visual inference, publication bias, small study effects

Webpages: https://CRAN.R-project.org/package=metaviz, https://metaviz.shinyapps.io/funnelinf_app/

Visual inference (Buja et al. 2009) is the formal inferential framework to test if graphically displayed data do or do not support a hypothesis. The general idea is that if the data supports an alternative hypothesis, the graphical display showing the real data should be identifiable when simultaneously presented with displays of simulated data under the null hypothesis. When compared to conventional statistical tests, visual inference showed promising results in experiments, for example, for testing linear model coefficients using boxplots and scatterplots (Majumder, Hofmann, and Cook 2013). With the package **nullabor** (Wickham, Chowdhury, and Cook 2014) helpful general purpose functions for visual inference are available within *R*. Due to the often uncertain or even misleading nature of funnel plot based conclusions, we identified funnel plots as a prime candidate field for the application of visual inference. For this purpose, we developed the function `funnelinf` which is available within the *R* package **metaviz**. The function `funnelinf` is specifically tailored to visual inference of funnel plots, for instance, with options for displaying significance contours, Egger's regression line, and for using different meta-analytic models for null plot simulation. In addition, the functionalities of `funnelinf` are made available as a shiny app for the convenient use by meta-analysts not familiar with *R*. Visual funnel plot inference and the capabilities of `funnelinf` are illustrated with real data from a meta-analysis on the mozart effect. Furthermore, results of an empirical experiment evaluating the power of visual funnel plot inference compared to traditional statistical funnel plot based tests are presented. Implications of these results are discussed and specific guidelines for the use of visual funnel plot inference are given.

References

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367: 4361–83.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108: 942–56.

Wickham, Hadley, Niladri Roy Chowdhury, and Di Cook. 2014. *Nullabor: Tools for Graphical Inference.* https://CRAN.R-project.org/package=nullabor.

useR! 2017 Oral Presentation

Integrated analysis of digital PCR experiments in R

Michał Burdukiewicz[1], *Piotr Sobczyk*[2], *Paweł Mackiewicz*[1], *Andrej-Nikolai Spiess*[3] *and Stefan Rödiger*[4]

1. University of Wrocław, Department of Genomics
2. Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics
3. University Medical Center Hamburg-Eppendorf
4. Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology

Keywords: digital PCR, multiple comparison, GUI, reproducible research

Webpages: dpcR package, dpcR website, pcRuniveRsum

Digital PCR (dPCR) is a variant of PCR, where the PCR amplification is conducted in multiple small volume reactions (termed partitions) instead of a bulk. The dichotomous status of each partition (positive or negative amplification) is used for absolute quantification of the template molecules by Poisson transformation of the proportion of positive partitions. The vast expansion of dPCR technology and its applications has been followed by the development of statistical data analysis methods. Yet, the software landscape is scattered, consisting of scripts in various programming languages, web servers with narrow scopes or closed source vendor software packages, that are usually tightly tied to their platform. This leads to unfavourable environments, as results from different platforms, or even from different laboratories using the same platform, cannot be easily compared with one another.

To address these challenges, we developed the dpcReport shiny server that provides an open-source tool for the analysis of dPCR data. dpcReport provides a streamlined analysis framework to the dPCR community that is compatible with the data output (e.g., CSV, XLSX) from different dPCR platforms (e.g., Bio-Rad QX100/200, Biomark). This goes beyond the basic dPCR data analysis with vendor-supplied softwares, which is often limited to the computation of the mean template copy number per partition and its uncertainty. dpcReport gives users more control over their data analysis and they benefit from standardization and reproducible analysis.

Our web server analyses data regardless of the platform vendor or type (droplet or chamber dPCR). It is not limited to the commercially available platforms and can also be used with experimental systems by importing data through the universal REDF format, which follows the IETF RFC 4180 standard. dpcReport provides users with advanced tools for data quality control and it incorporates statistical tests for comparing multiple reactions in an experiment (Burdukiewicz et al. 2016), currently absent in many dPCR-related software tools. dpcReport provides users with advanced tools for data quality control. The conducted analyses are fully integrated within extensive and customizable interactive HTML reports including figures, tables and calculations.

To improve reproducibility and transparency, a report may include snippets in *R* enabling an exact reproduction of the analysis performed by dpcReport. We developed **dpcR** package to collect all functionalities employed by the shiny server. Furthermore, the package provides additional functions facilitating analysis and quality control of dPCR data. Nevertheless, core functionalities are available through the shiny server to minimize entry barrier required to use our software.

Both dpcReport and *dpcR* follow the standardized dPCR nomenclature of the dMIQE guidelines (Huggett and Whale 2013). Since the vast functionality offered by our software may be overwhelming at first, our software is extensively documented. The documentation is enriched by the analysis of sample data sets.

References

Burdukiewicz, Michał, Stefan Rödiger, Piotr Sobczyk, Peter Schierack, and Paweł Mackiewicz. 2016. "Methods of Comparing Digital PCR Experiments." *Biomolecular Detection and Quantification* 28 (9): 14–19. doi:10.1016/j.bdq.2016.06.004.

Huggett, Jim F, and Alexandra Whale. 2013. "Digital PCR as a Novel Technology and Its Potential Implications for Molecular Diagnostics." *Clinical Chemistry* 59 (12): 1691–3. doi:10.1373/clinchem.2013.214742.

IRT test equating with the R package equateIRT

Michela Battauz[1]

1. Deparment of Economics and Statistics, University of Udine.

Keywords: Equating, Item Response Theory, Multiple Forms, Scoring, Testing.

Webpages: https://CRAN.R-project.org/package=equateIRT

In many testing programs, security reasons require that test forms are composed of different items, making test scores not comparable across different administrations. The equating process aims to provide comparable test scores. This talk focuses on Item Response Theory (IRT) methods for dichotomous items. In IRT models, the probability of a correct response depends on the latent trait under investigation and on the item parameters. Due to indentifiability issues, the latent variable is usually assumed to have zero mean and variance equal to one. Hence, when the model is fitted separately for different groups of examinees, the item parameter estimates are expressed on different measurement scales. The scale conversion can be achieved by applying a linear transformation of the item parameters, and the coefficients of this equation are called equating coefficients. This talk explains the functionalities of the *R* package **equateIRT** (Battauz 2015), which implements the estimation of the equating coefficients and the computation of the equated scores. Direct equating coefficients between pairs of forms that share some common items can be estimated using the mean-mean, mean-geometric mean, mean-sigma, Haebara and Stocking-Lord methods. However, the linkage plans are often quite complex, and not all forms can be linked directly. As proposed in Battauz (2013), the package computes also the indirect equating coefficients for a chain of forms and the average equating coefficients when two forms can be linked through more than one path. Using the equating coefficients so obtained, the item parameter estimates are converted to a common metric and it is possible to compute comparable scores. For this task, the package implements the true score equating and the observed score equating methods. Standard errors of the equating coefficients and the equated scores are also provided.

References

Battauz, Michela. 2013. "IRT Test Equating in Complex Linkage Plans." *Psychometrika* 78 (3): 464–80. doi:10.1007/s11336-012-9316-y.

———. 2015. "EquateIRT: An R Package for Irt Test Equating." *Journal of Statistical Software* 68 (1): 1–22. doi:10.18637/jss.v068.i07.

Deep Learning for Natural Language Processing in R

Miguel Fierro[1] and Angus Taylor[1]

1. Data Scientists at Microsoft

Keywords: Deep Learning, Natural Language Processing

Webpages: Blog | mxnet package | tutorial

The use of deep learning for NLP has attracted a lot of interest in the research community over recent years. This talk describes how deep learning techniques can be applied to natural language processing (NLP) tasks using *R*. We demonstrate how the **MXNet** deep learning framework can be used to implement, train and deploy deep neural networks that can solve text categorization and sentiment analysis problems.

We begin by briefly discussing the motivation and theory behind applying deep learning to NLP tasks. Deep learning has achieved a lot of success in the domain of image recognition. State-of-the-art image classification systems employ convolutional neural networks (CNNs) with a large number of layers. These networks perform well because they can learn hierarchical representations of the input with increasing levels of abstraction. In the context of NLP, neural networks have been shown to achieve good results. In particular, Recurrent Neural Networks such as Long Short Term Memory Networks (LSTMs) perform well for problems where the input is a sequence, such as speech recognition and text understanding. In this talk we explore an interesting approach which takes inspiration from the image recognition domain and applies CNNs to NLP problems. This is achieved by encoding segments of text in an image-like matrix, where each encoded word or character is equivalent to a pixel in the image.

CNNs have achieved excellent performance for text categorization and sentiment analysis. In this talk, we demonstrate how to implement a CNN for these tasks in *R*. As an example, we describe in detail the code to implement the Crepe model. To train this network, each input sentence is transformed into a matrix in which each column represents a one-hot encoding of each character. We describe the code needed to perform this transformation and how to specify the structure of the network and hyperparameters using the *R* bindings to *MXNet* provided in the **mxnet** package. We show how we implemented a custom *C++* iterator class to efficiently manage the input and output of data. This allows us to process CSV files in chunks, taking batches of raw text and tranforming them into matrices in memory, whilst distributing the computation over multiple GPUs. We describe how to set up a virtual machine with GPUs on Microsoft Azure to train the network, including installation of the necessary drivers and libraries. The network is trained on the Amazon categories dataset which consists of a training set of 2.38 million sentences, each of which map to one of 7 categories including *Books*, *Electronics* and *Home & Kitchen*.

The talk concludes with a demo of how a trained network can be deployed to classify new sentences. We demonstrate how this model can be deployed as a web service which can be consumed from a simple web app. The user can query the web service with a sentence and the API will return a product category. Finally, we show how the Crepe model can be applied to the sentiment analysis task using exactly the same network structure and training methods.

Through this talk, we aim to give the audience insight into the motivation for employing CNNs to solve NLP problems. Attendees will also gain an understanding of how they can be implemented, efficiently trained and deployed in *R*.

metawRite: Review, write and update meta-analysis results

Natalia da Silva[1], *Heike Hofmann*[1] *and Annette O'Connor*[1]
1. Iowa State University

Keywords: Living systematic review, meta-analysis, shiny, reproducible research

Webpages: https://github.com/natydasilva/metawRite

Systematic reviews are used to understand how treatments are effective and to design disease control policies, this approach is used by public health agencies such as the World Health Organization. Systematic reviews in the literature often include a meta-analysis that summarizes the findings of multiple studies. It is critical that such reviews are updated quickly as new scientific information becomes available, so the best evidence is used for treatment advice. However, the current peer-reviewed journal based approach to publishing systematic reviews means that reviews can rapidly become out of date and updating is often delayed by the publication model. Living systematic reviews have been proposed as a new approach to dealing with this problem. The main concept of a living review is to enable rapid updating of systematic reviews as new research becomes available, while also ensuring a transparent process and reproducible review. Our approach to a living systematic review will be implemented in an R package named **metawRite**. The goal is to combine writing and analysis of the review, allowing versioning and updating in an *R* package . **metawRite** package will allow an easy and effective way to display a living systematic review available in a web-based display. Three main tasks are needed to have an effective living systematic review: the ability to produce dynamic reports, availability online with an interface that enables end users to understand the data and the ability to efficiently update the review (and any meta-analysis) with new research (Elliott et al. 2014). **metawRite** package will cover these three task integrated in a friendly web based environment for the final user. This package is not a new meta-analysis package instead will be flexible enough to read different output models from the most used meta-analysis packages in *R* (**metafor** (Viechtbauer 2010), **meta** (Schwarzer 2007) among others), organize the information and display the results in an user driven interactive dashboard. The main function of this package will display a modern web-based application for update a living systematic review. This package combines the power of *R*, **shiny** (Chang et al. 2017) and **knitr** (Xie 2015) to get a dynamic reports and up to date meta-analysis results remaining user friendly. The package has the potential to be used by a large number of groups that conduct and update systematic review such as What Works clearinghouse (https://ies.ed.gov/ncee/WWC/) which reviews education interventions, Campbell Collaboration https://www.campbellcollaboration.org that includes reviews on topics such as social and criminal justice issues and many other social science topics, the Collaboration for Environment Evidence (http://www.environmentalevidence.org) and food production and security (http://www.syreaf.org) among others.

References

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2017. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Elliott, Julian H, Tari Turner, Ornella Clavisi, James Thomas, Julian PT Higgins, Chris Mavergames, and Russell L Gruen. 2014. "Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap." *PLoS Med* 11 (2). Public Library of Science: e1001603.

Schwarzer, Guido. 2007. "Meta: An R Package for Meta-Analysis." *R News* 7 (3): 40–45.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the metafor Package." *Journal of Statistical Software* 36 (3): 1–48. http://www.jstatsoft.org/v36/i03/.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. Vol. 29. CRC Press.

FFTrees: An R package to create, visualise and use fast and frugal decision trees

Phillips, Nathaniel D. [1], *Neth, Hansjoerg* [1,2], *Gaissmaier, Wolfgang* [2] *and Woike, Jan* [3]
1. University of Basel
2. University of Konstanz
3. Max Plank Institute for Human Development

Keywords: decision trees, decision making, package, visualization

Webpages: https://CRAN.R-project.org/package=FFTrees

Many complex real-world problems call for fast and accurate classification decisions. An emergency room physician faced with a patient complaining of chest pain needs to quickly decide if the patient is having a heart attack or not. A lost hiker, upon discovering a patch of mushrooms, needs to decide whether they are safe to eat or are poisonous. A stock portfolio adviser, upon seeing that, at 3:14 am, an influential figure tweeted about a 5 company he is heavily invested in, needs to decide whether to move his shares or sit tight. These decisions have important consequences and must be made under time-pressure with limited information. How can and should people make such decisions? One effective way is to use a fast and frugal decision tree (FFT). FFTs are simple heuristics that allow people to make fast, accurate decisions based on limited information (Gigerenzer and Goldstein 1996; Martignon, Katsikopoulos, and Woike 2008). In contrast to compensatory decision algorithms such as regression, or computationally intensive algorithms such as random forests, FFTs allow people to make fast decisions 'in the head' without requiring statistical training or a calculation device. Because they are so easy to implement, they are especially helpful in applied decision domains such as emergency rooms, where people need to be able to make decisions quickly and transparently.

While FFTs are easy to implement, actually constructing an effective FFT from data is less straightforward. While several FFT construction algorithms have been proposed 15 (Dhami and Ayton 2001; Martignon, Katsikopoulos, and Woike 2008; Martignon et al. 2003), none have been programmed and distributed in an easy-to-use and well-documented tool. The purpose of this paper is to fill this gap by introducing **FFTrees**, an *R* package that allows anyone to create, evaluate, and visualize FFTs from their own data. The package requires minimal coding, is documented by many examples, and provides quantitative performance measures and visual displays showing exactly how cases are classified at each level in the tree.

This presentation is structured in three sections: Section 1 provides a theoretical background on binary classification decision tasks and explains how FFTs solve them. Section 2 provides a 5-step tutorial on how to use the **FFTrees** package to construct and evaluate FFTs from data. Finally, Section 3 compares the prediction performance of **FFTrees** to alternative algorithms such as logistic regression and random forests. To preview our results, we find that trees created by **FFTrees** are both more efficient, and as accurate as the best of these algorithms across a wide variety of applied datasets. Moreover, they produce much simpler trees than standard decision tree algorithms such as **rpart** do, while maintaining similar prediction performance.

References

Dhami, Mandeep K, and Peter Ayton. 2001. "Bailing and Jailing the Fast and Frugal Way." *Journal of Behavioral Decision Making* 14 (2). Wiley Online Library: 141–68.

Gigerenzer, Gerd, and Daniel G Goldstein. 1996. "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review* 103 (4). American Psychological Association: 650.

Martignon, Laura, Konstantinos V Katsikopoulos, and Jan K Woike. 2008. "Categorization with Limited Resources: A Family of Simple Heuristics." *Journal of Mathematical Psychology* 52 (6). Elsevier: 352–61.

Martignon, Laura, Oliver Vitouch, Masanori Takezawa, and Malcolm R Forster. 2003. "Naive and yet Enlightened: From Natural Frequencies to Fast and Frugal Decision Trees." *Thinking: Psychological Perspective on Reasoning, Judgment, and Decision Making*, 189–211.

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

Nicholas Tierney[1], *Dianne Cook*[2], *Miles McBain*[3]

1. *Monash University, Department of econometrics and business statistics nicholas.tierney@gmail.com*
2. *Monash University, Department of econometrics and business statistics dicook@monash.edu*
3. *Queensland University of Technology, ARC Centre of Excellence for Statistical and Mathematical Frontiers milesmcbain@gmail.com*

Keywords: 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 Wickham (2009). Additionally, imputed missing data are not typically distinguished in visualisation and data summaries. Tidy data structures described in Wickham (2014) 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.

References

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (1): 1–23.

Letting R sense the world around it with **shinysense**

Nick Strayer[1] and Lucy D'Agostino McGowan[1], and Jeff Leek[2]

1. Vanderbilt University
2. Johns Hopkins University

Keywords: Shiny, JavaScript, Data Collection, User Experience

Webpages: https://github.com/nstrayer/shinysense, https://nickstrayer.shinyapps.io/shinysense_earr_demo/, https://nickstrayer.shinyapps.io/shinysense_swipr_demo/

shinysense is a package containing shiny modules all geared towards helping users make mobile-first apps for collecting data, or helping **Shiny** "sense" the outside world. Currently the package contains modules for gathering data on swiping (`shinyswipr`), audio (`shinyearr`), and from accelerometers (`shinymovr`). The goal of these functions is to take **Shiny** from a tool for demonstrating finished models or workflows into being a tool for data collection, enabling its use for training/testing models or building richer user experiences.

Several demo apps are contained in the package, including training and testing a speech recognition system using `shinyearr` and detecting spell casts performed by swinging your phone like a wand with `shinymovr`. In addition, the package is already being used in real-world products. Notable examples being the app *papr* which allows users to rapidly read and react to abstracts by swiping cards containing their content using `shinyswipr`, validating algorithm output in GenomeBot Tweet Generator, and contributr: an app that allows users to review github issues on various R packages.

A major goal of the construction of **shinysense** was mobile-friendly behavior. The massive proliferation of smartphones laden with sensors is a potential goldmine of data and use cases for statisticians and data scientists. This package attempts to help users harness this new flood of opportunities. A side effect of mobile oriented design is increased usability in non-static environments (Dou and Sundar 2016, Wigdor, Fletcher, and Morrison (2009)). For instance: an app running on a smartphone allowing physicians to input parameters into clinical models and instantly see the results (`shinyswipr`), to the generation and testing of fitness tracking algorithms by carrying a phone in a pocket (`shinymovr`).

Much in keeping with the primary goal of **Shiny**, by bringing powerful software tools such as inputs typically reserved for *JavaScript/* native apps to a tool used by scientists such as *R* we hope to lower the costs (monetary and otherwise) of bringing innovative applications to fruition.

References

Dou, Xue, and S Shyam Sundar. 2016. "Power of the Swipe: Why Mobile Websites Should Add Horizontal Swiping to Tapping, Clicking, and Scrolling Interaction Techniques." *International Journal of Human-Computer Interaction* 32 (4). Taylor & Francis: 352–62.

Wigdor, Daniel, Joe Fletcher, and Gerald Morrison. 2009. "Designing User Interfaces for Multi-Touch and Gesture Devices." In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, 2755–8. ACM.

An LLVM-based Compiler Toolkit for R

Nick Ulle

Department of Statistics, University of California, Davis

Keywords: compiler, code analysis, performance

Webpages: https://github.com/nick-ulle/rstatic, https://github.com/duncantl/Rllvm

R allows useRs to focus on the problems they want to solve by abstracting away the details of the hardware. This is a major contributor to *R*'s success as a data analysis language, but also makes *R* too slow and resource-hungry for certain tasks. Traditionally, useRs have worked around this limitation by rewriting bottlenecks in *Fortran*, *C*, or *C++*. These languages provide a substantial performance boost at the cost of abstraction, a trade-off that useRs should not have to face.

This talk introduces a collection of packages for analyzing, optimizing, and building compilers for *R* code, extending earlier work by Temple Lang (2014). By building on top of the *LLVM Compiler Infrastructure* (Lattner and Adve 2004), a mature open-source library for native code generation, these tools enable translation of *R* code to specialized machine code for a variety of hardware. Moreover, the tools are extensible and ease the creation of efficient domain-specific languages based on *R*, such as **nimble** and **dplyr**. Potential applications will be discussed and a simple compiler (inspired by *Numba*) for mathematical *R* code will be presented as a demonstration.

References

Lattner, Chris, and Vikram Adve. 2004. "LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation." In *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-Directed and Runtime Optimization*, 75. CGO '04. Washington, DC, USA: IEEE Computer Society.

Temple Lang, Duncan. 2014. "Enhancing R with Advanced Compilation Tools and Methods." *Statistical Science* 29 (2). Institute of Mathematical Statistics: 181–200.

ReinforcementLearning: A package for replicating human behavior in *R*

Nicolas Pröllochs, Stefan Feuerriegel and Dirk Neumann

University of Freiburg

Keywords: Reinforcement Learning, Human-Like Learning, Experience Replay, Q-Learning, Decision Analytics

Webpages: https://github.com/nproellochs/ReinforcementLearning

Reinforcement learning has recently gained a great deal of traction in studies that call for human-like learning. In settings where an explicit teacher is not available, this method teaches an agent via interaction with its environment without any supervision other than its own decision-making policy. In many cases, this approach appears quite natural by mimicking the fundamental way humans learn. However, implementing reinforcement learning is programmatically challenging, since it relies on continuous interactions between an agent and its environment. In fact, there is currently no package available that performs model-free reinforcement learning in *R*. As a remedy, we introduce the **ReinforcementLearning** *R* package, which allows an agent to learn optimal behavior based on sample experience consisting of states, actions and rewards. The result of the learning process is a highly interpretable reinforcement learning policy that defines the best possible action in each state. The package provides a remarkably flexible framework and is easily applied to a wide range of different problems. We demonstrate the added benefits of human-like learning using multiple real-world examples, e.g. by teaching the optimal movements of a robot in a grid map.

Differentiation of brain tumor tissue using hierarchical non-negative matrix factorization

Nicolas Sauwen
OpenAnalytics NV, Belgium

Keywords: non-negative matrix factorization, magnetic resonance imaging, brain tumor

Treatment of brain tumors is complicated by their high degree of heterogeneity. Various stages of the disease can occur throughout the same lesion, and transitions between the pathological tissue regions (i.e. active tumor, necrosis and edema) are diffuse (Price et al. 2006). Clinical practice could benefit from an accurate and reproducible method to differentiate brain tumor tissue based on medical imaging data.

We present a hierarchical variant of non-negative matrix factorization (hNMF) for characterizing brain tumors using multi-parametric magnetic resonance imaging (MRI) data (Sauwen et al. 2015). Non-negative matrix factorization (NMF) decomposes a non-negative matrix X into 2 factor matrices W and H, thereby providing a parts-based representation of the input data. In the current context, the columns of X correspond to the image voxels and the rows represent the different MRI parameters. The columns of W represent tissue-specific signatures and the rows of H contain the relative abundances per tissue type over the different voxels.

hNMF is available as an *R* package on CRAN and compatible with the **NMF** package. Besides the standard NMF algorithms that come with the **NMF** package, an effcient NMF algorithm called hierarchical alternating least-squares NMF was implemented and used within the hNMF framework. hNMF can be used as a general matrix factorization technique, but in the context of this talk it will be shown that valid tissue signatures are obtained using hNMF. Tissue abundances can be mapped back to the imaging domain, providing tissue differentiation on a voxel-wise basis (see Figure 1).

References

Price, SJ, R Jena, NG Burnet, PJ Hutchinson, AF Dean, A Pena, JD Pickard, TA Carpenter, and JH Gillard. 2006. “Improved Delineation of Glioma Margins and Regions of Infiltration with the Use of Diffusion Tensor Imaging: An Image-Guided Biopsy Study.” *American Journal of Neuroradiology* 27 (9). Am Soc Neuroradiology: 1969–74.

Sauwen, Nicolas, Diana M Sima, Sofie Van Cauter, Jelle Veraart, Alexander Leemans, Frederik Maes, Uwe Himmelreich, and Sabine Van Huffel. 2015. “Hierarchical Non-Negative Matrix Factorization to Characterize Brain Tumor Heterogeneity Using Multi-Parametric MRI.” *NMR in Biomedicine* 28 (12). Wiley Online Library: 1599–1624.

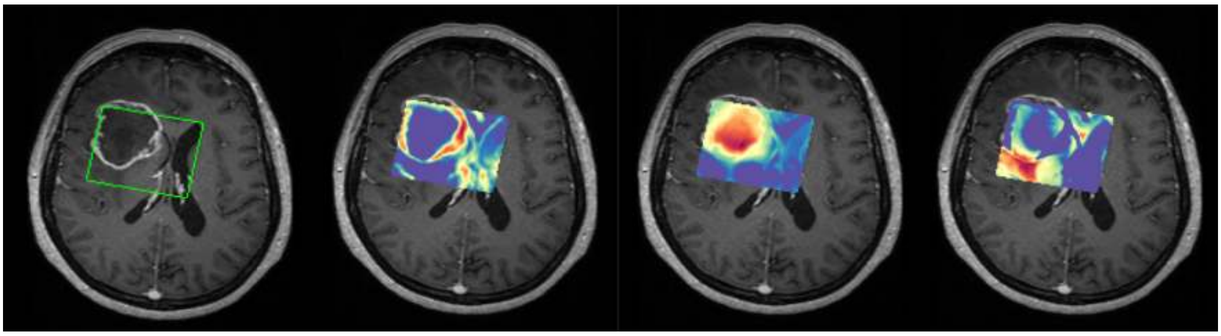


Figure 1: Figure 1: hNMF abundance maps of the pathological tissue regions of a glioblastoma patient. Left to right: T_1-weighted background image with region of interest (green frame); abundance map active tumor; abundance map necrosis; abundance map edema.

Show me the errors you didn't look for

Claus Thorn Ekstrøm[1] *and Anne Helby Petersen*[1]

1. Section of Biostatistics, University of Copenhagen

Keywords: Data cleaning, Quality control, Reproducible research, Data validation

Webpages: https://CRAN.R-project.org/package=dataMaid, https://github.com/ekstroem/dataMaid

The inability to replicate scientific studies has washed over many scientific fields in the last couple of years with potentially grave consequences. We need to give this problem its due diligence: Extreme care is needed when considering the representativeness of the data, and when we convey reproducible research information. We should not just document the statistical analyses and the data but also the exact steps that were part of the data cleaning process so we know which potential errors that we are unlikely to identify in the data.

Data cleaning and -validation are the first steps in any data analysis since the validity of the conclusions from the statistical analysis hinges on the quality of the input data. Mistakes in the data arise for any number of reasons, including erroneous codings, malfunctioning measurement equipment, and inconsistent data generation manuals. Ideally, a human investigator should go through each variable in the dataset and look for potential errors — both in input values and codings — but that process can be very time-consuming, expensive and error-prone in itself.

We present the *R* package **dataMaid** which implements an extensive and customizable suite of quality assessment tools to identify and document potential problems in the variables of a dataset. The results can be presented in an auto-generated, non-technical, stand-alone overview document intended to be perused by an investigator with an understanding of the variables in the dataset, but not necessarily knowledge of *R*. Thereby, **dataMaid** aids the dialogue between data analysts and field experts, while also providing easy documentation of reproducible data cleaning steps and data quality control. **dataMaid** also provides a suite of more typical *R* tools for interactive data quality assessment and -cleaning.

Depth and depth-based classification with *R*-package **ddalpha**

Oleksii Pokotylo[1], Pavlo Mozharovskyi[2], and Rainer Dyckerhoff[1]

1. University of Cologne
2. CREST, Ensai, Universite Bretagne Loire

Keywords: Data depth, Supervised classification, DD-plot, Outsiders, Visualization

Webpages: https://cran.r-project.org/package=ddalpha

Following the seminal idea of John W. Tukey, data depth is a function that measures how close an arbitrary point of the space is located to an implicitly defined center of a data cloud. Having undergone theoretical and computational developments, it is now employed in numerous applications with classification being the most popular one. The *R*-package **ddalpha** is a software directed to fuse experience of the applicant with recent achievements in the area of data depth and depth-based classification.

ddalpha provides an implementation for exact and approximate computation of most reasonable and widely applied notions of data depth. These can be further used in the depth-based multivariate and functional classifiers implemented in the package, where the $DD\alpha$-procedure is in the main focus. The package is expandable with user-defined custom depth methods and separators. The implemented functions for depth visualization and the built-in benchmark procedures may also serve to provide insights into the geometry of the data and the quality of pattern recognition.

How we built a Shiny App for 700 users?

Olga Mierzwa-Sulima[1], Marek Rogala[1], Filip Stachura[1]

1. Appsilon Data Science

Keywords: Shiny, shiny.semantic, UI, UX, application performance, analytics consulting

Shiny has proved itself a great tool for communicating data science teams' results. However, developing a Shiny app for a large scope project that will be used commercially by more than dozens of users is not easy. The first challenge is User Interface (UI): the expectations are that the app should not vary from modern web pages. Secondly, performance directly impacts user experience (UX), and it's difficult to maintain efficiency with growing requirements and user base.

In this talk, we will share our experience from a real-life case study of building an app used daily by 700 users where our data science team tackled all these problems. This, to our knowledge, was one of the biggest production deployments of a Shiny App.

We will show an innovative approach to building a beautiful and flexible Shiny UI using **shiny.semantic** package (an alternative to standard Bootstrap). Furthermore, we will talk about the non-standard optimization tricks we implemented to gain performance. Then we will discuss challenges regarding complex reactivity and offer solutions. We will go through implementation and deployment process of the app using a load balancer. Finally, we will present the application and give details on how this benefited our client.

Beyond Prototyping: Best practices for R in critical enterprise environments

Oliver Bracht

eoda GmbH

Keywords: data science, business, industry, best practice, critical enterprise environments, R,

Over the last couple of years, R has become increasingly popular among business users. Today, it is the first choice of many data science departments when it comes to ad-hoc analysis and data visualization, research and prototyping.

But when it comes to critical production environments, IT departments are still reluctant to consider R as part of their software stack. And there are reasons for that: Dynamic typing, the reputation for being slow (still around!), the lack of experience regarding management and administration of R (and its 10,000 packages), to name some of them.

Nevertheless, with the help of some friends, it is feasible and reasonable to use R as a core application even in critical production environments. This talk will share lessons learned from practical experience and point out a best practice landscape of tools, approaches and methods to make that happen.

GNU R on a Programmable Logic Controller (PLC) in an Embedded-Linux Environment

Oliver Glaß and Philipp Luchscheider
1. ZAE Bayern - Germany

Keywords: *GNU R*, PLC, Embedded, Linux

Webpages: ZAE Bayern e.V. | Wago PLC 750-8208 | Package httpuv | Package ZMQ

Being one of the leading institutions in the field of applied energy research, the Bavarian Center for Applied Energy Research (ZAE Bayern) combines excellent research with excellent economic implementation of the results. Our main research goal is to increase the capacity of low-voltage grids for installed photovoltaics. Therefore the influences were analysed by taking measurements in grid nodes and households. In this context we applied several open-source programming languages, but we found *GNU R* to be best suiting. This results from being capable of analysing, manipulating and plotting the measurement data as well as simulating and controlling our real test systems. We have installed multiple storages and modules in different households. In order to control these test sites, we use Wago PLC 750-8202, that are currently programmed in *Codesys*. As strategies can get quite complex (due to individual forecasting, dynamic non-linear storage models, . . .), we see that *Codesys* isn't capable of this.

With choosing *R* for complex computations, we have access to a wide rage of libraries and our self-developed strategies used for analysis and simulation relying on the measurement data from the grid and the weather. To bring the strategy to the PLC, we divided the whole system into two parts. One of it is running on our central servers and is preparing external data from our databases for each test site. Therefore, we set up a control platform using *node-red* and the **httpuv** package in order to run *R* scripts on demand. The second half will be computed on the Wago PLCs. With the board support package (BSP), Wago provides a tool-chain to its customers for build their own customised firmware. Our proposed idea is to get *R* and *Python* together with the basic *Codesys* running on the PLC. As for that, *Python* will serve as an asynchronous local controller, that is able to start calculations in *R* and provide control quantities to *Codesys*. We try to apply the **rzmq** package for inter process communication (IPC) and data exchange. For example, the information delivered to the *Python* controller will be cyclically forwarded to the `global.environment` of a continuously running *R* instance. This helps us to reduce the start-up and initialization effort for our models. On demand, *R* is instructed to calculate a given strategy, that is chosen for a specific day and situation by the central servers. *R* will hand the result back to the *Python* controller and forward it to *Codesys*, where short-term closed control-loops can be established. Our approach will be tested and verified on our Wago PLCs in our environment.

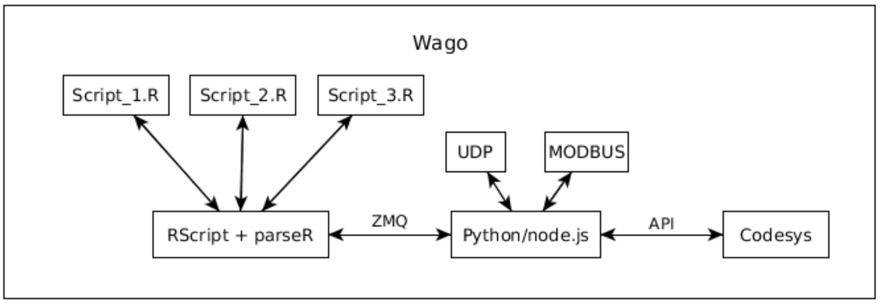


Figure 1: Schematic procedure inside the PLC

Teaching psychometrics and analysing educational tests with **ShinyItemAnalysis**

Patrícia Martinková[1], Adéla Drabinová[1,2] and Jakub Houdek[1,3]

1. *Institute of Computer Science, Czech Academy of Sciences, Prague*
2. *Faculty of Mathematics and Physics, Charles University, Prague*
3. *Faculty of Informatics and Statistics, University of Economics, Prague*

Keywords: psychometrics, educational test, item response theory, shiny, R

Webpages: https://CRAN.R-project.org/package=ShinyItemAnalysis, https://shiny.cs.cas.cz/ShinyItemAnalysis/

This work introduces **ShinyItemAnalysis** (Martinková, Drabinová, Leder, & Houdek, 2017) *R* package and an online shiny application for psychometric analysis of educational tests and their items.

ShinyItemAnalysis covers broad range of methods and offers data examples, model equations, parameter estimates, interpretation of results, together with selected *R* code, and is thus suitable for teaching psychometric concepts with *R*. It is based on examples developed for course of Item Response Theory models for graduate students at University of Washington.

Besides, the application aspires to be a simple tool for analysis of educational tests by allowing the users to upload and analyze their own data and to automatically generate analysis report in *PDF* or *HTML*. It has been used at workshops for educators developing admission tests and in development of instruments for classroom testing such as concept inventories, see McFarland et al. (2017).

We argue that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of admission test to medical school we demonstrate how **ShinyItemAnalysis** may provide a simple and free tool to routinely analyze tests and to explain advanced psychometric models to students and those who develop educational tests.

References

Martinková, P., Drabinová, A., Leder, O., & Houdek, J. (2017). *ShinyItemAnalysis: Test and item analysis via shiny.* Retrieved from shiny.cs.cas.cz/ShinyItemAnalysis/; https://CRAN.R-project.org/package=ShinyItemAnalysis

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., & Modell, H. (2017). Development and validation of the Homeostasis Concept Inventory. *CBE-Life Sciences Education.*

useR! 2017 Oral Presentation

Urban green spaces and their biophonic soundscape component

Paul Devos

Department of Information Technology, Ghent University

Keywords: soundscape ecology, urbanization, green space, indicators, soundscape

Abstract

Sustainable urban environments with urban green spaces like city parks and urban gardens provide enduring benefits for individuals and society. Providing recreational spaces they encourage physical activity resulting in improved physical and mental health of citizens. As such, the density and the quality of these areas are of high importance in urban area planning.

In order to study urban green spaces as a landscape, the study of their soundscape as the holistic experience of their sounds has recently gained attention in soundscape ecological studies. Using *R*, the **soundecology** and **seewave** packages provide accessible processing tools appropriate to automate the calculation of soundecology indicators of long run sound recordings from permanent outdoor recorders. These indicators give information about the biophonic component in the present soundscape, and as such give a clear indication of the quality of the green space. Since bird vocalizations contribute strongly to the biophonic component, their spring singing activity is clearly reflected in the yearly pattern of these indicators.

A pilot study focussing on the annual variations of the soundscape of a typical urban green space has been conducted.

codebookr: Codebooks in *R*

Peter Baker[1]

1. *Senior Statistical Consultant and Senior Lecturer*

Keywords: code book, data dictionary, data cleaning, validation, automation

Webpages: https://github.com/petebaker/codebookr, https://github.com/ropensci/auunconf/issues/46

codebookr is an *R* package under development to automate cleaning, checking and formatting data using metadata from Codebooks or Data Dictionaries. It is primarily aimed at epidemiological research and medical studies but can be easily used in other research areas.

Researchers collecting primary, secondary or tertiary data from RCTs or government and hospital administrative systems often have different data documentation and data cleaning needs to those scraping data off the web or collecting in-house data for business analytics. However, all studies will benefit from using codebooks which comprehensively document all study variables including derived variables. Codebooks document data formats, variable names, variable labels, factor levels, valid ranges for continuous variables, details of measuring instruments and so on.

For statistical consultants, each new data set has a new codebook. While statisticians may get a photocopied codebook or pdf, my preference is a spreadsheet so that the metadata can be used directly. Many data analysts are happy to use this metadata to code syntax to read, clean and check data. I prefer to automate this process by reading the codebook into *R* and then using the metadata directly for data checking, cleaning, factor level definitions.

While there is considerable interest in the data wrangling and cleaning (Jonge and Loo 2013; Wickham 2014; Fischetti 2017), there appear to be few tools available to read codebooks (see http://jason.bryer.org/posts/2013-01-10/Function_for_Reading_Codebooks_in_R.html) and even less to automatically apply the metadata to datasets.

We outline the fundamentals of **codebookr** and demonstrate it's use on examples of research projects undertaken at University of Queensland's School of Public Health.

References

Fischetti, Tony. 2017. *Assertr: Assertive Programming for R Analysis Pipelines*. https://CRAN.R-project.org/package=assertr.

Jonge, Edwin de, and Mark van der Loo. 2013. "An Introduction to Data Cleaning with R." Technical Report 201313. Statistics Netherlands. http://cran.vinastat.com/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf.

Wickham, Hadley. 2014. "Tidy Data." *The Journal of Statistical Software* 59 (10). http://www.jstatsoft.org/v59/i10/.

Show Me Your Model: tools for visualisation of statistical models

Przemysław Biecek[1,2]

1. *University of Warsaw*
2. *Warsaw University of Technology*

Keywords: Model visualisation, model exploration, structure visualisation, grammar of model visualisation

The **ggplot2** (Wickham 2009) package changed the way how we approach to data visualisation. Instead of looking for suitable type of a plot out of dozens of predefined templates now we express the relation among variables with a well defined grammar based on the excellent book *The Grammar of Graphics* (Wilkinson 2006).

Similar revolution is happening with tools for visualisation of statistical models. In the *CRAN* repository, one may find a lot of great packages that graphically explain a structure or diagnostic for some family of statistical models. Just to mention few known and powerful packages: **rms**, **forestmodel** and **regtools** (regression models), **survminer** (survival models), **ggRandomForests** (random forest based models), **factoextra** (multivariate structure exploration), **factorMerger** (one-way ANOVA) and many, many others. They are great, but they do not share same logic nor structure.

New packages from the *tidyverse*, like **broom** (Robinson 2017), creates an opportunity to build an unified interface for model exploration and visualisation for large collection of statistical models. And there is more and more articles that set theoretical foundations for unified grammar of model visualization (see for example Wickham, Cook, and Hofmann 2015).

In this talk I am going to present various approaches to the model visualisation, give an overview of selected existing packages for visualisation of statistical models and discuss proposition for a unified grammar of model visualisation.

References

Robinson, David. 2017. *Broom: Convert Statistical Analysis Objects into Tidy Data Frames.* https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. *Visualizing Statistical Models: Removing the Blindfold.* Statistical Analysis; Data Mining 8(4).

Wilkinson, Leland. 2006. *The Grammar of Graphics.* Springer Science & Business Media.

shadow: R Package for Geometric Shade Calculations in an Urban Environment

Michael Dorman, Adi Vulkan, Evyatar Erell, Itai Kloog
Ben-Gurion University of the Negev

Keywords: shadow, sun position, geometry, solar radiation, building facades

Webpage: https://CRAN.R-project.org/package=shadow

Spatial analysis of the urban environment frequently requires estimating whether a given point is shaded or not, given a representation of spatial obstacles (e.g. buildings) and a time-stamp with its associated solar position. For example, we may be interested in -

- Calculating the amount of time a given roof or facade is shaded, to determine the utility of installing Photo-Voltaic cells for electricity production.
- Calculating shade footprint on vegetated areas, to determine the expected microclimatic influence of a new tall building.

These types of calculations are usually applied in either vector-based 3D (e.g. ESRI's ArcScene) or raster-based 2.5D (i.e. Digital Elevation Model, DEM) settings. However, the former solutions are mostly restricted to proprietary software associated with specific 3D geometric model formats. The latter DEM-based solutions are more common, in both open-source (e.g. GRASS GIS) as well as proprietary (e.g. ArcGIS) software. The **insol** R package provides such capabilities in R. Though conceptually and technically simpler to work with, DEM-based approaches are less suitable for an urban environment, as opposed to natural terrain, for two reasons -

- A continuous elevation surface at sufficiently high resolution for the urban context (e.g. LIDAR) may not be available and is expensive to produce.
- DEMs cannot adequately represent individual vertical urban elements (e.g. building facades), thus limiting the interpretability of results.

The **shadow** package aims at addressing these limitations. Functions in this package operate on a vector layer of building outlines along with their heights (class `SpatialPolygonsDataFrame` from package **sp**), rather than a DEM. Such data are widely available, either from local municipalities or from global datasets such as OpenStreetMap. Currently functions to calculate shadow height, Sky View Factor (SVF) and shade footprint on ground are implemented. Since the inputs are vector-based, the resulting shadow estimates are easily associated with specific urban elements such as buildings, roofs or facades.

We present a case study where package **shadow** was used to calculate shading on roofs and facades in a large neighborhood (Rishon-Le-Zion city, Israel), on an hourly temporal resolution and a 1-m spatial resolution. The results were combined with Typical Meteorological Year (TMY) direct solar radiation data to derive total annual insolation for each 1-m grid cell. Subsequently the locations where installation of photovoltaic (PV) cells is worthwhile, given a predefined threshold production, were mapped.

The approach is currently applicable to a flat terrain and does not treat obstacles (e.g. trees) other than the buildings. Our case study demonstrates that subject to these limitations package **shadow** can be used to calculate shade and insolation estimates in an urban environment using widely available polygonal building data. Future development of the package will be aimed at combining vector-associated shadow calculations with raster data representing non-flat terrain.

useR! 2017 Oral Presentation

kmlShape: clustering longitudinal data according to their shapes

René Echochard [1,2,3]
Christophe Genolini [4,5]
1. Hospices Civils de Lyon, Service de Biostatistique-Bioinformatique, Lyon, France.
2. Universit? de Lyon, Lyon, France.
3. CNRS, UMR 5558, Laboratoire de Biom?trie et Biologie ?volutive, ?quipe Biostatistique-Sant?, F-69100, Villeurbanne France.
4. Universit? Paris-Nanterre
5. Zebrys

Keywords: Clustring, Longitudinal data, k-means, dynamic time warping, Frechet distance

Longitudinal data are data in which each variable is measured repeatedly over time. One possibility for the analysis of such data is to cluster them. The majority of clustering methods group together individual that have close trajectories at given time points. These methods group trajectories that are locally close but not necessarily those that have similar shapes. However, in several circumstances, the progress of a phenomenon may be more important than the moment at which it occurs. One would thus like to achieve a partitioning where each group gathers individuals whose trajectories have similar shapes whatever the time lag between them.

In this presentation, we introduce a longitudinal data partitioning algorithm based on the shapes of the trajectories rather than on classical distances. Because this algorithm is time consuming, we propose as well two data simplification procedures that make it applicable to high dimensional datasets.

In an application to Alzheimer disease, this algorithm revealed a "rapid decline" patient group that was not found by the classical methods. In another application to the feminine menstrual cycle, the algorithm showed, contrarily to the current literature, that the luteinizing hormone presents two peaks in an important proportion of women (22%).

Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure

Rianne Schouten[1], Peter Lugtig[1] and Gerko Vink[1]

1. University of Utrecht, Department of Methodology and Statistics

Keywords: R-function `ampute`, Multivariate Amputation, Missing Data Methodology, Simulation Studies

Webpages: https://github.com/RianneSchouten/mice/blob/ampute/vignettes/Vignette_Ampute.pdf, https://www.rdocumentation.org/packages/mice/versions/2.30/topics/ampute, https://cran.r-project.org/web/packages/mice/index.html

Abstract: Missing data are a ubiquitous problem in scientific research, especially since most statistical analyses require complete data. To evaluate the performance of methods dealing with missing data, researchers perform simulation studies. An important aspect of these studies is the generation of missing values in complete data (i.e. the amputation procedure) and this procedure will be our focus.

Since no amputation software was available, we developed and implemented an extensive amputation procedure into an R-function: `ampute` (available in multiple imputation package **mice**). We will show that the multivariate amputation approach generates legitimate missing data problems.

NA

We will provide evidence that `ampute` overcomes the problems of stepwise univariate amputation. With `ampute`, we have an efficient amputation method to accurately evaluate missing data methodology.

rTRNG: Advanced Parallel Random Number Generation in R

Riccardo Porreca[1] and Roland Schmid[1]

1. Mirai Solutions GmbH

Keywords: Random Number Generation, Monte Carlo, Parallel Execution, Reproducibility

Webpages: https://github.com/miraisolutions/rTRNG

Monte Carlo simulations provide a powerful computational approach to address a wide variety of problems in several domains, such as physical sciences, engineering, computational biology and finance. The independent-samples and large-scale nature of Monte Carlo simulations make the corresponding computation suited for parallel execution, at least in theory. In practice, pseudo-random number generators (RNGs) are intrinsically sequential. This often prevents having a parallel Monte Carlo algorithm that is *playing fair*, meaning that results are independent of the architecture, parallelization techniques and number of parallel processes (Mertens 2009; Bauke 2016).

We will show that parallel-oriented RNGs and techniques in fact exist and can be used in R with the **rTRNG** package (Porreca, Schmid, and Bauke 2017). The package relies on TRNG (Bauke 2016), a state-of-the-art *C++* pseudo-random number generator library for sequential and parallel Monte Carlo simulations.
TRNG provides *parallel* RNGs that can be manipulated by `jumping` ahead an arbitrary number of steps or `splitting` a sequence into any desired subsequence(s), thus supporting techniques such as block-splitting and leapfrogging suitable to parallel algorithms.

The **rTRNG** package provides access to the functionality of the underlying TRNG *C++* library by embedding its sources and headers. Beyond this, it makes use of **Rcpp** and **RcppParallel** to offer several ways of creating and manipulating pseudo-random streams, and drawing random variates from them, which we will demonstrate:

- Base-*R*-like usage for selecting and manipulating the current engine, as a simple and immediate way for *R* users to use **rTRNG**
- *Reference objects* wrapping the underlying *C++* TRNG random number engines can be created and manipulated in OOP-style, for greater flexibility in using parallel RNGs in *R*
- TRNG *C++* library and headers can be accessed directly from within *R* projects that use *C++*, both via standalone *C++* code (via `sourceCpp`) or through creating an *R* package that depends on **rTRNG**

References

Bauke, Heiko. 2016. *Tina's Random Number Generator Library*. https://numbercrunch.de/trng/trng.pdf.

Mertens, Stephan. 2009. "Random Number Generators: A Survival Guide for Large Scale Simulations." In *Modern Computational Science 09*. BIS-Verlag.

Porreca, Riccardo, Roland Schmid, and Heiko Bauke. 2017. *rTRNG: R Package Providing Access and Examples to TRNG C++ Library*. https://github.com/miraisolutions/rTRNG/.

Computer Vision and Image Recognition algorithms for R users

Jan Wijffels - BNOSAC - www.bnosac.be

Keywords: Computer Vision, Image recognition, Object detection, Image feature engineering

Webpages: https://github.com/bnosac/image

R has already quite some packages for image processing, namely magick, imager, EBImage and OpenImageR.

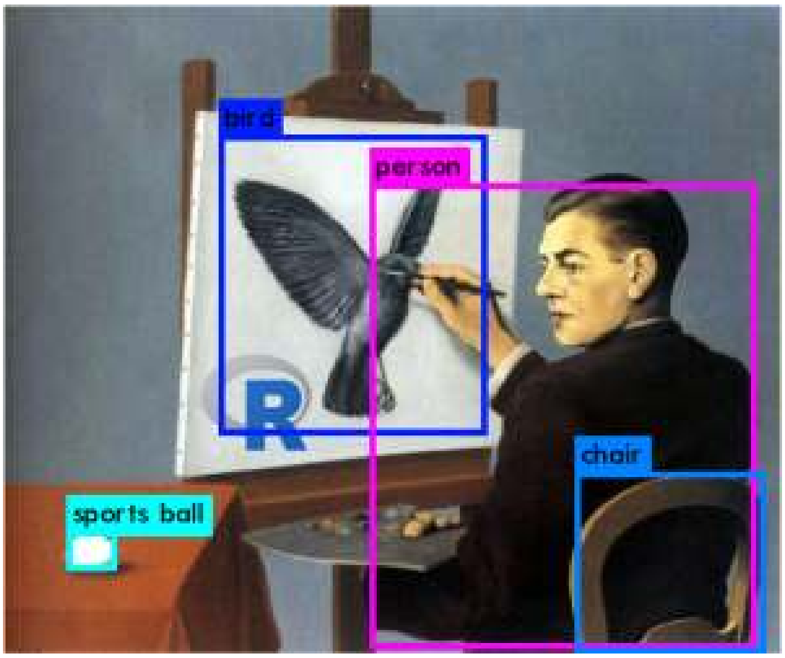
The field of image processing is rapidly evolving with new algorithms and techniques quickly popping up from Learning and Detection, to Denoising, Segmentation and Edges, Image Comparison and Deep Learning.

In order to complement these existing packages with new algorithms, we implemented a number of *R* packages. Some of these packages have been released on https://github.com/bnosac/image, namely:

- **image.CornerDetectionF9**: FAST-9 corner detection for images (license: BSD-2).
- **image.LineSegmentDetector**: Line Segment Detector (LSD) for images (license: AGPL-3).
- **image.ContourDetector**: Unsupervised Smooth Contour Line Detection for images (license: AGPL-3).
- **image.CannyEdges**: Canny Edge Detector for Images (license: GPL-3).
- **image.dlib**: Speeded up robust features (SURF) and histogram of oriented gradients (HOG) features (license: AGPL-3).
- **image.darknet**: Image classification using darknet with deep learning models AlexNet, Darknet, VGG-16, Extraction (GoogleNet) and Darknet19. As well object detection using the state-of-the art YOLO detection system (license: MIT).
- **dlib**: dlib: Allow Access to the 'Dlib' C++ Library (license: BSL-1.0)

More packages and extensions will be released in due course.

In this talk, we provide an overview of these newly developed packages and the new computer vision algorithms made accessible for R users.



Robets: Forecasting with Robust Exponential Smoothing with Trend and Seasonality

Ruben Crevits[1] and Christophe Croux[1]

1. KU Leuven

Keywords: Time Series, Forecasting, Robust Statistics, Exponential Smoothing

Webpages: https://CRAN.R-project.org/package=robets, https://rcrevits.wordpress.com/

Simple forecasting methods, such as exponential smoothing, are very popular in business analytics. This is not only due to their simplicity, but also because they perform very well, in particular for shorter time series. Incorporating trend and seasonality into an exponential smoothing method is standard. Many real time series, show seasonal patterns that should be exploited for forecasting purposes. Including a trend or not may be less clear. For instance, weekly sales (in units) may show an increasing trend, but the sales will not grow to infinity. Here, the damped trend model gives an outcome. Damped trend exponential smoothing gives excellent results in forecasting competitions.

In a highly cited paper, Hyndman and Khandakar (2008) developed an automatic forecasting method using exponential smoothing, available as the *R* package **forecast**. We propose the package **robets**, an outlier robust alternative of the function `ets` in the **forecast** package. For each method of a class of exponential smoothing variants we made a robust alternative. The class includes methods with a damped trend and/or seasonal components. The robust method is developed by robustifying every aspect of the original exponential smoothing variant. We provide robust forecasting equations, robust initial values, robust smoothing parameter estimation and a robust information criterion. The method is an extension of Gelper, Fried, and Croux (2010) and is described in more detail in Crevits and Croux (2016).

The code of the developed *R* package is based on the function `ets` of the **forecast** package. The usual functions for visualizing the models and forecasts also work for `robets` objects. Additionally there is a function `plotOutliers` which highlights outlying values in a time series.

References

Crevits, Ruben, and Christophe Croux. 2016. “Forecasting with Robust Exponential Smoothing with Damped Trend and Seasonal Components.” *Working Paper.*

Gelper, S, R Fried, and C Croux. 2010. “Robust Forecasting with Exponential and Holt-Winters Smoothing.” *Journal of Forecasting* 29: 285–300.

Hyndman, R J, and Y Khandakar. 2008. “Automatic Time Series Forecasting: The Forecast Package for R.” *Journal of Statistical Software* 27 (3).

IntegratedJM - an R package to Jointly Model the Gene-Expression and Bioassay Data, Taking Care of the Fingerprint Feature effect

Rudradev Sengupta[1]*, Nolen Joy Perualila*[1] *and Ziv Shkedy*[1]

1. Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, 3590 Diepenbeek, Belgium

Keywords: Bioactivity, Biomarkers, Chemical Structure, Joint Model, Multi-source

Webpages: https://cran.r-project.org/web/packages/IntegratedJM/index.html

In recent days, data from different sources need to be integrated together in order to arrive at meaningful conclusions. In drug-discovery experiments, most of the different data sources, related to a new set of compounds under development, are of high-dimension. For example, in order to investigate the properties of a new set of compounds, pharmaceutical companies need to analyse chemical structure (fingerprint features) of the compounds, phenotypic bioactivity (bioassay read-outs) data for targets of interest and transcriptomic(gene expression) data. Perualila-Tan et al. (2016) proposed a joint model in which the three data sources are included to better detect the association between gene expression and biological activity. For a given set of compounds, the joint modeling approach accounts for a possible effect of the chemical structure of the compound on both variables. The joint model allows us to identify genes as potential biomarkers for compound's efficacy. The joint modeling approach, proposed by Perualila-Tan et al. (2016), is implemented in the **IntegratedJM** R package which provides, in addition to model estimation and inference, a set of exploratory and visualization functions that can be used to clearly present the results. The joint model and the **IntegratedJM** R package are discussed in details in Perualila et al. (2016) as well.

References

Perualila, Nolen Joy, Ziv Shkedy, Rudradev Sengupta, Theophile Bigirumurame, Luc Bijnens, Willem Talloen, Bie Verbist, Hinrich W.H. Göohlmann, Adetayo Kasim, and QSTAR Consortium. 2016. "Applied Surrogate Endpoint Evaluation Methods with Sas and R." In, edited by Ariel Alonso, Theophile Bigirumurame, Tomasz Burzykowski, Marc Buyse, Geert Molenberghs, Leacky Muchene, Nolen Joy Perualila, Ziv Shkedy, and Wim Van der Elst, 275–309. CRC Press.

Perualila-Tan, Nolen, Adetayo Kasim, Willem Talloen, Bie Verbist, Hinrich W.H. Göhlmann, QSTAR Consortium, and Ziv Shkedy. 2016. "A Joint Modeling Approach for Uncovering Associations Between Gene Expression, Bioactivity and Chemical Structure in Early Drug Discovery to Guide Lead Selection and Genomic Biomarker Development." *Statistical Applications in Genetics and Molecular Biology* 15: 291–304. doi:10.1515/sagmb-2014-0086.

A Tidy Data Model for Natural Language Processing

Taylor Arnold[1]

1. Department of Math and Computer Science, University of Richmond

Keywords: text analysis, database normalization, exploratory data analysis, feature extraction, visualization

Webpages: https://CRAN.R-project.org/package=cleanNLP, https://github.com/statsmaths/cleanNLP

This talk introduces the R package **cleanNLP**, which provides a set of fast tools for converting a textual corpus into a set of normalized tables. The underlying natural language processing pipeline utilizes Stanford's CoreNLP library, exposing a number of annotation tasks for text written in English, French, German, and Spanish (Marneffe et al. 2016, De Marneffe et al. (2014)). Annotators include tokenization, part of speech tagging, named entity recognition, entity linking, sentiment analysis, dependency parsing, coreference resolution, and information extraction (Lee et al. 2011). The functionality provided by the package applies the tidy data philosophy (Wickham 2014) to the processing of raw textual data by offering three distinct contributions:

- a data schema representing the output of an NLP annotation pipeline as a collection of normalized tables;
- a set of native Java output functions converting a Stanford CoreNLP annotation object directly, without converting into an intermediate XML format, into this collection of normalized tables;
- tools for converting from the tidy model into (sparse) data matrices appropriate for exploratory and predictive modeling.

Together, these contributions simplify the process of doing exploratory data analysis over a corpus of text. The output works seamlessly with both tidy data tools as well as other programming and graphing systems. The talk will illustrate the basic usage of the **cleanNLP** package, explain the rational behind the underlying data model, and show an example from a corpus of the text from every State of the Union address made by a United States President (Peters 2016).

References

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. "Universal Stanford Dependencies: A Cross-Linguistic Typology." In *LREC*, 14:4585–92.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. "Stanford's Multi-Pass Sieve Coreference Resolution System at the Conll-2011 Shared Task." In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 28–34. Association for Computational Linguistics.

Marneffe, Marie de, Filip Ginter, Yoav Goldberg, Jan Hajic, Chris Manning, Ryan McDonald, Joakim Nivre, et al. 2016. "Universal Dependencies."

Peters, Gerhard. 2016. "State of the Union Addresses and Messages."

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics.

EpiModel: An R Package for Mathematical Modeling of Infectious Disease over Networks

Samuel M. Jenness,[1] Steven M. Goodreau,[2] and Martina Morris[3]
1. Department of Epidemiology, Emory University
2. Department of Anthropology, University of Washington
3. Departments of Statistics and Sociology, University of Washington

Keywords: mathematical model, infectious disease, epidemiology, networks, R

Webpages: https://CRAN.R-project.org/package=EpiModel, http://epimodel.org/

The **EpiModel** package provides tools for building, simulating, and analyzing mathematical models for epidemics using *R*. Epidemic models are a formal representation of the complex systems that collectively determine the population dynamics of infectious disease transmission: contact between people, inter-host infection transmission, intra-host disease progression, and the underlying demographic processes. Simulating epidemic models serves as a computational laboratory to gain insight into the dynamics of these disease systems, test empirical hypotheses about the determinants of a specific outbreak patterns, and forecast the impact of interventions like vaccines, clinical treatment, or public health education campaigns.

A range of different modeling frameworks has been developed in the field of mathematical epidemiology over the last century. Several of these are included in **EpiModel**, but the unique contribution of this software package is a general stochastic framework for modeling the spread of epidemics across dynamic contact networks. Network models represent repeated contacts with the same person or persons over time (e.g., sexual partnerships). These repeated contacts give rise to persistent network configurations – pairs, triples, and larger connected components – that in turn may establish the temporally ordered pathways for infectious disease transmission across a population. The timing and sequence of contacts, and the transmission acts within them, is most important when transmission requires intimate contact, that contact is relatively rare, and the probability of infection per contact is relatively low. This is the case for HIV and other sexually transmitted infections.

Both the estimation and simulation of the dynamic networks in **EpiModel** are implemented using Markov Chain Monte Carlo (MCMC) algorithm functions for exponential-random graph models (ERGMs) from the **statnet** suite of *R* packages. These MCMC algorithms exploit a key property of ERGMs: that the maximum likelihood estimates of the model parameters uniquely reproduce the model statistics in expectation. The mathematical simulation of the contact network over time is theoretically guaranteed to vary stochastically around the observed network statistics. Temporal ERGMs provide the only integrated, principled framework for both the estimation of dynamic network models from sampled empirical data and also the simulation of complex dynamic networks with theoretically justified methods for handling changes in population size and composition over time.

In this talk, I will provide an overview of both the modeling tools built into **EpiModel**, designed to facilitate learning for students new to modeling, and the package's application programming interface (API) for extending **EpiModel**, designed to facilitate the exploration of novel research questions for advanced modelers. I will motivate these research-level extensions by discussing our recent applications of these network modeling statistical methods and software tools to investigate the transmission dynamics of HIV and sexually transmitted infections among men who have sex with men in the United States and heterosexual couples in Sub-Saharan Africa.

Performance Benchmarking of the R Programming Environment on Knight's Landing

Scott Michael[1] *and James McCombs*[1]
1. Indiana University, Pervasive Technology Institute

Keywords: Multicore architectures, benchmarking, scalability, Xeon Phi

We present performance results obtained with a new performance benchmark of the R programming environment on the Xeon Phi Knights Landing (KNL) and standard Xeon-based compute nodes. The benchmark package consists of microbenchmarks of matrix linear algebra kernels and machine learning functionality included in the *R* distribution that can be built from those kernels. Our microbenchmarking results show that the KNL compute nodes exhibited similar or superior performance compared to the standard Xeon-based nodes for matrix dimensions of moderate to large size for most of the microbenchmarks, executing as much as five times faster than the standard Xeon-based nodes. For the clustering and neural network training microbenchmarks, the standard Xeon-based nodes performed up to four times faster than their Xeon Phi counterparts for many large data sets, indicating that commonly used *R* packages may need to be reengineered to take advantage of existing optimized, scalable kernels.

Over the past several years a trend of increased demand for high performance computing (HPC) in data analysis has emerged. This trend is driven by increasing data sizes and computational complexity(Fox et al. 2015; Kouzes et al. 2009). Many data analysts, researchers, and scientists are turning to HPC machines to help with algorithms and tools, such as machine learning, that are computationally demanding and require large amounts of memory (Raj et al. 2015). The characteristics of large scale machines (e.g. large amounts of RAM per node, high storage capacity, and advanced processing capabilities) appear very attractive to these researchers, however, challenges remain for algorithms to make optimal use of the hardware (Lee et al. 2014). Depending on the nature of the analysis to be performed, analytics workflows may be carried out as many independent concurrent processes requiring little or no coordination between them, or as highly coordinated parallel processes in which the processes perform portions of the same computational task. Regardless of the implementation, it is important for data analysts to have software environments at their disposal which can exploit the performance advantages of modern HPC machines.

We developed an R performance benchmark to determine the single-node run time performance of compute intensive linear algebra kernels that are common to many data analytics algorithms, and the run time performance of machine learning functionality commonly implemented with linear algebra operations. We then performed single-node strong scaling tests of the benchmark on both Xeon and Xeon Phi based systems to determine problem sizes and numbers of threads for which the KNL architecture was comparable to or outperformed their standard Intel Xeon counterparts. It is our intention that these results be used to guide future performance optimization efforts of the *R* programming environment to increase the applicability of HPC machines for compute-intensive data analysis. The benchmark is also generally applicable to a variety of systems and architectures and can be easily run to determine the computational potential of a system when using *R* for many data analysis tasks.

References

Fox, Geoffrey, Judy Qiu, Shantenu Jha, Saliya Ekanayake, and Supun Kamburugamuve. 2015. "Big Data, Simulations and Hpc Convergence." In *Workshop on Big Data Benchmarks*, 3–17. Springer.

Kouzes, Richard T, Gordon A Anderson, Stephen T Elbert, Ian Gorton, and Deborah K Gracio. 2009. "The Changing Paradigm of Data-Intensive Computing." *Computer* 42 (1). IEEE: 26–34.

Lee, Seunghak, Jin Kyu Kim, Xun Zheng, Qirong Ho, Garth A Gibson, and Eric P Xing. 2014. "On Model Parallelization and Scheduling Strategies for Distributed Machine Learning." In *Advances in Neural Information Processing Systems*, 27:2834–42.

Raj, Pethuru, Anupama Raman, Dhivya Nagaraj, and Siddhartha Duggirala. 2015. "High-Performance Big-Data Analytics." *Computing Systems and Approaches (Springer, 2015)* 1. Springer.

An Efficient Algorithm for Solving Large Fixed Effects OLS Problems with Clustered Standard Error Estimation

Thomas Balmat[1] and Jerome Reiter[1]

1. Duke University

Keywords: large data least squares, fixed effects estimation, clustered standard error estimation, sparse matrix methods, high performance computing

Large fixed effects regression problems, involving order 10^7 observations and 10^3 effects levels, present special computational challenges but, also, a special performance opportunity because of the large proportion of entries in the expanded design matrix (fixed effect levels translated from single columns into dichotomous indicator columns, one for each level) that are zero. For many problems, the proportion of zero entries is above 0.99995, which would be considered sparse. In this presentation, we demonstrate an efficient method for solving large, sparse fixed effects OLS problems without creation of the expanded design matrix and avoiding computations involving zero-level effects. This leads to minimal memory usage and optimal execution time. A feature, often desired in social science applications, is to estimate parameter standard errors clustered about a key identifier, such as employee ID. For large problems, with ID counts in the millions, this presents a significant computational challenge. We present a sparse matrix indexing algorithm that produces clustered standard error estimates that, for large fixed effects problems, is many times more efficient than standard "sandwich" matrix operations.

Social contact data in endemic-epidemic models and probabilistic forecasting with **surveillance**

Sebastian Meyer[1] and Johannes Bracher[2] and Leonhard Held[2]

1. *Friedrich-Alexander-Universität Erlangen-Nürnberg*
2. *Universität Zürich*

Keywords: age-structured contact matrix, areal count time series, infectious disease epidemiology, norovirus, spatio-temporal surveillance data

Webpages: https://CRAN.R-project.org/package=surveillance

Routine surveillance of notifiable infectious diseases gives rise to weekly counts of reported cases stratified by region and age group. A well-established approach to the statistical analysis of such surveillance data are endemic-epidemic time-series models (`hhh4`) as implemented in the *R* package **surveillance** (Meyer, Held, and Höhle 2017). Autoregressive model components reflect the temporal dependence inherent to communicable diseases. Spatial dynamics are largely driven by human travel and can be captured by movement network data or a parametric power law based on the adjacency matrix of the regions. Furthermore, the social phenomenon of "like seeks like" produces characteristic contact patterns between subgroups of a population, in particular with respect to age (Mossong et al. 2008). We thus incorporate an age-structured contact matrix in the `hhh4` modelling framework to

1. assess age-specific disease spread while accounting for its spatial pattern (Meyer and Held 2017)
2. improve probabilistic forecasts of infectious disease spread (Held, Meyer, and Bracher 2017)

We analyze weekly surveillance counts on norovirus gastroenteritis from the 12 city districts of Berlin, in six age groups, from week 2011/27 to week 2015/26. The following year (2015/27 to 2016/26) is used to assess the quality of the predictions.

References

Held, Leonhard, Sebastian Meyer, and Johannes Bracher. 2017. "Probabilistic Forecasting in Infectious Disease Epidemiology: The Thirteenth Armitage Lecture." *bioRxiv*. doi:10.1101/104000.

Meyer, Sebastian, and Leonhard Held. 2017. "Incorporating Social Contact Data in Spatio-Temporal Models for Infectious Disease Spread." *Biostatistics* 18 (2): 338–51. doi:10.1093/biostatistics/kxw051.

Meyer, Sebastian, Leonhard Held, and Michael Höhle. 2017. "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package **surveillance**." *Journal of Statistical Software*. http://arxiv.org/abs/1411.0416.

Mossong, Joël, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, et al. 2008. "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *PLoS Medicine* 5 (3): e74. doi:10.1371/journal.pmed.0050074.

Modules in R

Sebastian Warnholz[1]

1. INWT Statistics GmbH

Keywords: programming, functional-programming

Webpages: https://CRAN.R-project.org/package=modules, https://github.com/wahani/modules

In this talk I present the concept of modules inside the *R* language. The key idea of the **modules** package is to provide a unit of source code which is self contained, i.e. has it's own scope. The main and most reliable infrastructure for such organisational units of source code is a package. Compared to a package modules can be considered ad-hoc, but still self contained. That means they come with a mechanism to import dependencies and also to export member functions. However, modules do not act as replacements for packages. Instead they are a unit of abstraction in between functions and packages.

There are two use cases in which modules can be beneficial. First when we write scripts and want to use sourced functions or, in general, need more control of the enclosing environment of a function. Here we may be interested to be able to state the dependencies of a function close to its definition; and also without the typical side effects to the *global environment* of the current *R* session. Second, as an organisational unit inside packages. Here modules can act as similar entities as objects in object-oriented-programming. However, other languages with similar concepts are mostly functional and the design borrows from languages like *julia*, *Erlang* and *F#*. As a result modules are not designed to contain data. Furthermore there is no formal mechanism for inheritance. Instead several possibilities for module composition are implemented.

R and Haskell: Combining the best of two worlds

Sigrid Keydana

Trivadis GmbH, München

Keywords: HaskellR, Haskell, Interoperability

Surely there's no need to explain to useR! attendees what's so great about *R*! *Haskell*, on the other hand, is a great language too - statically typed, purely functional, lazy, fast, and with that, you know, cool and mathy touch . . . ;-)

Statistics, data science, and machine learning, however, are not that easy to do from *Haskell*, as it does not have all the specialized and comfortable-to-use libraries *R* has.

Fortunately, the guys at tweag.io developed **HaskellR**, providing *Haskell* with full *R* interoperability. With **ihaskell-inline-R**, *R* can even be used in *IHaskell* notebooks. In this session, we'll show how to get started with **HaskellR**, and how you can get the best of both worlds.

How to deal with Missing Data in Time Series and the imputeTS package

Steffen Moritz[1] and Thomas Bartz-Beielstein[1]

1. TH Köln University of Applied Sciences

Keywords: Missing Data, Time Series, Imputation, Visualization, Data Pre-Processing

Webpages: https://CRAN.R-project.org/package=imputeTS, https://github.com/SteffenMoritz/imputeTS

In almost every domain from industry, finance, up to biology time series data is measured. One common problem that can come along with time series measurement are missing observations. During several projects with industry partners in the last years, we often experienced sensor malfunctions or transmission issues leading to missing sensor data. As subsequent processes or analysis methods may require missing values to be replaced with reasonable values up-front, missing data handling can be crucial.

This talk gives a short overview about methods for missing data in time series in *R* in general and subsequently introduces the **imputeTS** package. The **imputeTS** package is specifically made for handling missing data in time series and offers several functions for visualization and replacement (imputation) of missing data. Based on usage examples it is shown how **imputeTS** can be used for time series imputation.

Most well-known and established packages (e.g. **mice**, **VIM**, **AMELIA**, **missMDA**) for missing value imputation focus mostly on cross-sectional data, while methods for time series data are not that familiar to users. Also, from an algorithmic perspective, these two imputation use cases are slightly different: imputation for cross-sectional data relies on inter-attribute correlations, while (univariate) time series imputation needs to employ time dependencies. Overall, this talk is supposed to give users a first glance at time series imputation in *R* with special focus on the **imputeTS** package.

Implementing Predictive Analytics projects in corporate environments

Steffen Wagner[1] and Verena Pflieger[1]

1. INWT Statistics GmbH

Keywords: R in production, business applications

INWT Statistics is a company specialised on services around Predictive Analytics. For our clients we develop customised algorithms and solutions. While *R* is the de facto standard within our company, we face many challenges in our day to day work when we implement these solutions for our clients. To overcome these challenges we use standardised approaches for integrating predictive models into the infrastructure of our clients.

In this talk we present an overview of a typical project structure and the role of the *R* language within our projects. *R* is used as an Analytics tool, for automatic reporting, building dashboards, and various programming tasks. While developing solutions, we always have to keep in mind how our clients plan to utilise the results. Here we have experience with the full delivery of the outcome in the form of *R* packages and workshops, as well as giving access to the results by using dashboards or automatically generated reports. Different companies need different models of implementation. Thus in each project we have to decide early on how *R* can be used to its full potential to meet our clients requirements. In this regard, we give insights into various models of implementation and our experience with each of them.

Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in *R*

Susanne Berger[1], Nathaniel Graham[2] and Achim Zeileis[1]

1. *University of Innsbruck*
2. *Trinity University Texas*

Keywords: clustered data, clustered covariance matrix estimators, object-orientation, simulation, *R*

Webpages: http://R-forge.R-project.org/projects/sandwich/

Clustered covariances or clustered standard errors are very widely used to account for correlated or clustered data, especially in economics, political sciences, or other social sciences. They are employed to adjust the inference following estimation of a standard least-squares regression or generalized linear model estimated by maximum likelihood. Although many publications just refer to "the" clustered standard errors, there is a surprisingly wide variation in clustered covariances, particularly due to different flavors of bias corrections. Furthermore, while the linear regression model is certainly the most important application case, the same strategies can be employed in more general models (e.g. for zero-inflated, censored, or limited responses).

In *R*, the **sandwich** package (Zeileis 2004; Zeileis 2006) provides an object-oriented approach to "robust" covariance matrix estimation based on methods for two generic functions (`estfun()` and `bread()`). Using this infrastructure, sandwich covariances for cross-section or time series data have been available for models beyond `lm()` or `glm()`, e.g., for packages **MASS**, **pscl**, **countreg**, **betareg**, among many others. However, corresponding functions for clustered or panel data have been somewhat scattered or available only for certain modeling functions. This shortcoming has been corrected in the development version of **sandwich** on R-Forge. Here, we introduce this new object-oriented implementation of clustered and panel covariances and assess the methods' performance in a simulation study.

References

Zeileis, Achim. 2004. "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software* 11 (10): 1–17. http://www.jstatsoft.org/v11/i10/.

———. 2006. "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software* 16 (9): 1–16. http://www.jstatsoft.org/v16/i09/.

Automatically archiving reproducible studies with Docker

Daniel Nüst and Matthias Hinz

Institute for Geoinformatics, University of Münster, Germany

Keywords: Docker, Reproducible Research, Open Science

Webpage: https://github.com/o2r-project/containerit/

Reproducibility of computations is crucial in an era where data is born digital and analysed algorithmically. Most studies however only publish the results, often with figures as important interpreted outputs. But where do these figures come from? Scholarly articles must provide not only a description of the work but be accompanied by data and software. *R* offers excellent tools to create reproducible works, i.e. Sweave and RMarkdown. Several approaches to capture the workspace environment in *R* have been made, working around CRAN's deliberate choice not to provide explicit versioning of packages and their dependencies. They preserve a collection of packages locally (**packrat**, **pkgsnap**, **switchr**/**GRANBase**) or remotely (*MRAN timemachine*/**checkpoint**), or install specific versions from CRAN or source (**requireGitHub**, **devtools**). Installers for old versions of *R* are archived on CRAN. A user can manually re-create a specific environment, but this is a cumbersome task.

We introduce a new possibility to preserve a runtime environment including both, packages and *R*, by adding an abstraction layer in the form of a container, which can execute a script or run an interactive session. The package **containeRit** automatically creates such containers based on *Docker*. *Docker* is a solution for packaging an application and its dependencies, but shows to be useful in the context of reproducible research (Boettiger 2015). The package creates a container manifest, the `Dockerfile`, which is usually written by hand, from `sessionInfo()`, *R* scripts, or RMarkdown documents. The `Dockerfiles` use the *Rocker* community images as *base images*. Docker can build an executable image from a `Dockerfile`. The image is executable anywhere a Docker runtime is present. **containeRit** uses **harbor** for building images and running containers, and **sysreqs** for installing system dependencies of *R* packages. Before the planned CRAN release we want to share our work, discuss open challenges such as handling linked libraries (see discussion on geospatial libraries in Rocker), and welcome community feedback.

containeRit is developed within the DFG-funded project Opening Reproducible Research to support the creation of Executable Research Compendia (ERC) (Nüst et al. 2017).

References

Boettiger, Carl. 2015. "An Introduction to Docker for Reproducible Research, with Examples from the R Environment." *ACM SIGOPS Operating Systems Review* 49 (January): 71–79. doi:10.1145/2723872.2723882.

Nüst, Daniel, Markus Konkol, Edzer Pebesma, Christian Kray, Marc Schutzeichel, Holger Przibytzin, and Jörg Lorenz. 2017. "Opening the Publication Process with Executable Research Compendia." *D-Lib Magazine* 23 (January). doi:10.1045/january2017-nuest.

The Revised Sequential Parameter Optimization Toolbox

Thomas Bartz-Beielstein[1] and Martin Zaefferer[1] and Jörg Stork[1] and Sebastian Krey[1]
1. TH Köln University of Applied Sciences

Keywords: optimization, tuning, surrogate model, computer experiments

Webpages: https://CRAN.R-project.org/package=SPOT

Real-world optimization problems often have very high complexity, due to multi-modality, constraints, noise or other crucial problem features. For solving these optimization problems a large collection of methods are available. Most of these methods require to set a number of parameters, which have a significant impact on the optimization performance. Hence, a lot of experience and knowledge about the problem is necessary to give the best possible results. This situation grows worse if the optimization algorithm faces the additional difficulty of strong restrictions on resources, especially time, money or number of experiments.

Sequential parameter optimization (Bartz-Beielstein, Lasarczyk, and Preuss 2005) is a heuristic combining classical and modern statistical techniques for the purpose of efficient optimization. It can be applied:

- to efficiently tune and select the parameters of other search algorithms, or
- to optimize expensive-to-evaluate problems directly, shifting evaluations to a surrogate model.

SPO is especially useful in scenarios where

(1) no experience of how to choose the parameter setting of an algorithm is available,
(2) a comparison with other algorithms is needed,
(3) an optimization algorithm has to be applied effectively and efficiently to a complex problem, and
(4) the objective function is a black-box and expensive to evaluate.

The Sequential Parameter Optimization Toolbox **SPOT** provides enhanced statistical techniques such as design and analysis of computer experiments, different methods for surrogate modeling and optimization to effectively use sequential parameter optimization in the above mentioned scenarios.

Version 2 of the **SPOT** package is a complete redesign and rewrite of the original *R* package. Most function interfaces were redesigned to give a more streamlined usage experience. At the same time, modular and transparent code structures allow for increased extensibility. In addition, some new developments were added to the **SPOT** package. A Kriging model implementation, based on earlier Matlab code by Forrester et al. (Forrester, Sobester, and Keane 2008), has been extended to allow for the usage of categorical inputs. Additionally, it is now possible to use stacking for the construction of ensemble learners (Bartz-Beielstein and Zaefferer 2017). This allows for the creation of models with a far higher predictive performance, by combining the strengths of different modeling approaches.

In this presentation we show how the new interface of **SPOT** can be used to efficiently optimize the geometry of an industrial dust filter (cyclone). Based on a simplified simulation of this real world industry problem, some of the core features of **SPOT** are demonstrated.

References

Bartz-Beielstein, Thomas, and Martin Zaefferer. 2017. "Model-Based Methods for Continuous and Discrete Global Optimization." *Applied Soft Computing* 55: 154–67. doi:10.1016/j.asoc.2017.01.039.

Bartz-Beielstein, Thomas, Christian Lasarczyk, and Mike Preuss. 2005. "Sequential Parameter Optimization." In *Proceedings Congress on Evolutionary Computation 2005 (Cec'05)*, 1553. Edinburgh, Scotland. http://www.spotseven.de/wp-content/papercite-data/pdf/blp05.pdf.

Forrester, Alexander, Andras Sobester, and Andy Keane. 2008. *Engineering Design via Surrogate Modelling*. Wiley.

bradio: Add data music widgets to your business intelligence dashboards.

Thomas Levine

Keywords: music, sonification, Shiny

Webpages: http://src.thomaslevine.com/bradio/, https://thomaslevine.com/!/data-music/

Recent years have brought considerable advances in data sonification (Ligges et al. 2016; Sueur, Aubin, and Simonis 2008; Stone and Garisson 2012; Stone and Garrison 2013; Levine 2015), but data sonification is still a very involved process with many technical limitations. Developing data music in *R* has historically been a very tedious process because of *R*'s poor concurrency features and general weakness in audio rendering capabilities (Levine 2016). End-user data music tools can be more straightforward, but they usually constrain users to very particular and rudimentary aesthetic mappings (Siegert and Williams 2017; Levine 2014; Borum Consulting 2014). Finally, existing data music implementations have limited interactivity capabilities, and no integrated solutions are available for embedding in business intelligence dashboards.

I have addressed these various issues by implementing **bradio**, a *Shiny* widget for rendering data music. In **bradio**, a song is encoded as a *Javascript* function that can take data inputs from *R*, through *Shiny*. The *Javascript* component relies on the **webaudio** *Javascript* package (johnnyscript 2014) and is thus compatible with songs written for the **webaudio** *Javascript* package, the **baudio** *Javascript* package (substack 2014), and *Javascript* **code-music-studio** (substack 2015); this compatibility allows for existing songs to be adapted easily as data music. **bradio** merges the convenience of interactive *Javascript* music with the data analysis power of *R*, facilitating the prototyping and presentation of sophisticated interactive data music.

Borum Consulting. 2014. "Readme for Tonesintune Version 2.0." http://tonesintune.com/Readme.php.

johnnyscript. 2014. *webaudio*. 2.0.0 ed. https://www.npmjs.com/package/webaudio.

Levine, Thomas. 2014. *Sheetmusic*. 0.0.4 ed. https://pypi.python.org/pypi/sheetmusic.

———. 2015. "Plotting Data as Music Videos in R." In *UseR!* http://user2015.math.aau.dk/contributed_talks#61.

———. 2016. "Approaches to Live Music Synthesis for Multivariate Data Analysis in R." In *SatRday*. http://budapest.satrdays.org/.

Ligges, Uwe, Sebastian Krey, Olaf Mersmann, and Sarah Schnackenberg. 2016. *tuneR: Analysis of Music*. http://r-forge.r-project.org/projects/tuner/.

Siegert, Stefan, and Robin Williams. 2017. *Sonify: Data Sonification - Turning Data into Sound*. https://CRAN.R-project.org/package=sonify.

Stone, Eric, and Jesse Garisson. 2012. "Give Your Data a Listen." In *UseR!* http://biostat.mc.vanderbilt.edu/wiki/pub/Main/UseR-2012/81-Stone.pdf.

Stone, Eric, and Jesse Garrison. 2013. *AudiolyzR: Give Your Data a Listen*. https://CRAN.R-project.org/package=audiolyzR.

substack. 2014. *baudio*. 2.1.2 ed. https://www.npmjs.com/package/baudio.

———. 2015. *code-music-studio*. 1.5.2 ed. https://www.npmjs.com/package/code-music-studio.

———. n.d. "Make Music with Algorithms!" http://studio.substack.net/-/help.

Sueur, J., T. Aubin, and C. Simonis. 2008. "Seewave: A Free Modular Tool for Sound Analysis and Synthesis." *Bioacoustics* 18: 213–26. http://isyeb.mnhn.fr/IMG/pdf/sueuretal_bioacoustics_2008.pdf.

Maximum growth rate estimation with **growthrates**

Thomas Petzoldt[1], David Kneis[1,2] and Claudia Seiler[1,2]

1. TU Dresden, Institute of Hydrobiology, 01062 Dresden, Germany
2. Helmholtz-Centre for Environmental Research – UFZ, 39114 Magdeburg, Germany

Keywords: population growth, nonlinear models, differential equation

Webpages: https://CRAN.R-project.org/package=growthrates, https://github.com/tpetzoldt/growthrates

The population growth rate is a direct measure of fitness, common in many disciplines of theoretical and applied biology, e.g. physiology, ecology, eco-toxicology or pharmacology. The *R* package **growthrates** aims to streamline growth rate estimation from direct or indirect measures of population density (e.g. cell counts, optical density or fluorescence) of batch experiments or field observations. It can be applicable to different species of bacteria, protists, and metazoa, e.g. *E. coli*, *Cyanobacteria*, *Paramecium*, green algae or *Daphnia*.

The package includes three types of methods:

1. Fitting of linear models to the period of exponential growth using the "growth rates made easy"-method of Hall and Barlow (2013),
2. Nonparametric growthrate estimation by using smoothers. The current implementation uses function `smooth.spline`, similar to method of package **grofit** (Kahm et al. 2010),
3. Nonlinear fitting of parametric models like logistic, Gompertz, Baranyi or Huang (Huang 2011) is done with package **FME** (Flexible Modelling Environment) of Soetaert and Petzoldt (2010). Growth models can be given either in closed form or as numerically integrated system of differential equations, that are numerically solved with package **deSolve** (Soetaert, Petzoldt, and Setzer 2010) and **cOde** (Kaschek 2016).

The package contains methods to fit single data sets or complete experimental series. It uses S4 classes and contains functions for extracting results (e.g. `coef`, `summary`, `residuals`, ...), and methods for convenient plotting. The fits and the growth models can be visualized with **shiny** apps.

References

Hall, Acar, B. G., and M. Barlow. 2013. "Growth Rates Made Easy." *Mol. Biol. Evol.* 31: 232–38. doi:10.1093/molbev/mst197.

Huang, Lihan. 2011. "A New Mechanistic Growth Model for Simultaneous Determination of Lag Phase Duration and Exponential Growth Rate and a New Belehdredek-Type Model for Evaluating the Effect of Temperature on Growth Rate." *Food Microbiology* 28 (4): 770–76. doi:10.1016/j.fm.2010.05.019.

Kahm, Matthias, Guido Hasenbrink, Hella Lichtenberg-Frate, Jost Ludwig, and Maik Kschischo. 2010. "Grofit: Fitting Biological Growth Curves with R." *Journal of Statistical Software* 33 (7): 1–21. doi:10.18637/jss.v033.i07.

Kaschek, Daniel. 2016. *cOde: Automated C Code Generation for Use with the deSolve and bvpSolve Packages*. https://CRAN.R-project.org/package=cOde.

Soetaert, Karline, and Thomas Petzoldt. 2010. "Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME." *Journal of Statistical Software* 33 (3): 1–28. doi:10.18637/jss.v033.i03.

Soetaert, Karline, Thomas Petzoldt, and R. Woodrow Setzer. 2010. "Solving Differential Equations in R: Package deSolve." *Journal of Statistical Software* 33 (9): 1–25. doi:10.18637/jss.v033.i09.

useR! 2017 Oral Presentation

mapedit - interactive manipulation of spatial objects

Tim Appelhans[1] *and Kenton Russell*[2]

1. GfK Geomarketing | Nürnberg | Germany
2. TimelyPortfolio

Keywords: Spatial analysis, Interactive, Visualization

Webpages: https://github.com/r-spatial/mapedit, http://r-spatial.org/r/2017/01/30/mapedit_intro.html

The *R* ecosystem offers a powerful set of packages for geospatial analysis. For a comprehensive list see the CRAN Task View: Analysis of Spatial Data. Yet, many geospatial workflows require interactivity for smooth uninterrupted completion. This interactivity is currently restricted to viewing and visual inspection (e.g. packages **leaflet** and **mapview**) and, with very few exceptions, there is currently no way to manipulate spatial data in an interactive manner in *R*. One noteworthy exception is function `drawExtent` in the **raster** package which lets the user select a geographic sub-region of a given `Raster*` object on a static plot of the visualized layer and saves the resultant extent or subset in a new object (if desired). Such operations are standard spatial tasks and are part of all standard spatial toolboxes. With new tools, such as **htmlwidgets**, **shiny**, and **crosstalk**, we can now inject this useful interactivity without leaving the R environment.

Package **mapedit** aims to provide a set of tools for basic, yet useful manipulation of spatial objects within the *R* environment. More specifically, we will provide functionality to:

1. draw, edit and delete a set of new features on a blank map canvas,
2. edit and delete existing features,
3. select and query from a set of existing features,
4. edit attributes of existing features.

In this talk we will outline the conceptual and technical approach we take in **mapedit** to provide the above functionality and will provide a short live demonstration hightlighting the use of the package.

The **mapedit** project is being realized with financial support from the RConsortium.

The growing popularity of R in data journalism

Timo Grossenbacher[1]

1. SRF Data, Swiss Public Broadcast (http://www.srf.ch/data)

Keywords: Data Journalism, Journalism, Reproducibility, Automation, Reporting

Webpages: https://rddj.info,

https://timogrossenbacher.ch/2015/12/why-data-journalists-should-start-using-r-in-2016/

In this talk, Timo Grossenbacher, data journalist at Swiss Public Broadcast and creator of Rddj.info, will show that R is becoming more and more popular among a new community: data journalists. He will showcase some innovative work that has been done with R in the field of data journalism, both by his own team and by other media outlets all over the world. At the same time, he will point out the strengths (reproducibility, for example) and hurdles (having to learn to code) of using R for a typical data journalism workflow – a workflow that is often centered around quick, exploratory data analysis rather than statisticial modeling. During the talk, he will also point out and controversially discuss packages that are of great help for journalists especially, such as the **tidyverse**, **readxl** and **googlesheets** packages.

ShinyProxy

Tobias Verbeke[1]

1. Open Analytics NV

Keywords: Shiny, enterprise computing, open source

Webpages: https://shinyproxy.io

Shiny is nice technology to write interactive R-based applications. It has been rapidly adopted and the R community has collaborated on many interesting extensions. Until recently, though, deployments in larger organizations and companies required proprietary solutions. ShinyProxy fills this gap and offers a **fully open source** alternative to run and manage shiny applications at large.

In this talk we detail the ShinyProxy architecture and demonstrate how it meets the needs of organizations. First of all, by design ShinyProxy scales to thousands of **concurrent users**. Secondly, it offers **authentication and authorization** functionality using standard technologies like LDAP, ActiveDirectory, OpenID Connect as well as social authentication (Facebook, Twitter, Google, LinkedIn or Github). Thirdly, the management interface allows to **monitor application usage** real-time and provides infrastructure to collect usage statistics in event logging databases (e.g. influxdb) or databases for scientific computing (e.g. MonetDB). Finally, the ShinyProxy developers took special care to develop a solution that can be easily **embedded in broader applications** and (responsive) web sites.

Besides these basic features, the use of Docker technology opens a new world of possibilities that go beyond the current proprietary platforms and in the final section of the talk we will show how academic institutions, governmental organizations and industry roll out Shiny apps with ShinyProxy and, last but not least, how you can do this too.

useR! 2017 Oral Presentation

Taking Advantage of the Byte Code Compiler

Tomas Kalibera[1] *and Luke Tierney*[2]

1. Czech Technical University
2. University of Iowa

Keywords: Compiler, Byte-Code, Interpreter, Performance

Webpages: http://www.stat.uiowa.edu/~luke/R/compiler/compiler.pdf

Since version 2.13, R includes a byte-code compiler and interpreter which complement the older abstract-syntax-tree (AST) interpreter. The AST interpreter directly executes R code represented as a tree of objects produced by the parser. The byte-code compiler compiles the AST into a sequence of byte-code instructions, which is then interpreted using a byte-code interpreter. The byte-code compiler and interpreter implement a number of performance optimizations which speed up scalar code and particularly loops operating on scalar variables. Performance of highly vectorized code is likely to be unaffected by the byte-code compiler/interpreter as in fact such code spends most of the time executing outside the R interpreters. The compiler can compile packages at installation time and individual R functions on request. It can also transparently compile loops and functions as they execute (just-in-time).

The talk will show examples of code that runs particularly fast with the compiler, code that is unaffected by it, and code that runs particularly slow. The slowdowns are almost always due to time spent in compilation; once compiled and loaded, the code should not run slower than in the unoptimized AST interpreter. When triggered just-in-time, the compiler includes heuristics that try to prevent compilation in case it is not likely to pay off, but sometimes they are wrong. The talk will show on concrete examples how these overheads can be avoided. The talk will be technical and will be aimed at package authors and R users who write performance critical R code.

Transformation Forests

Torsten Hothorn[1] and Achim Zeileis[2]

1. University of Zurich, Zurich, Switzerland
2. University of Innsbruck, Innsbruck, Austria

Keywords: random forest, transformation model, quantile regression forest, conditional distribution, conditional quantiles

Webpages: https://R-forge.R-project.org/projects/ctm https://arxiv.org/1701.02110

Regression models for supervised learning problems with a continuous target are commonly understood as models for the conditional mean of the target given predictors. This notion is simple and therefore appealing for interpretation and visualisation. Information about the whole underlying conditional distribution is, however, not available from these models. A more general understanding of regression models as models for conditional distributions allows much broader inference from such models, for example the computation of prediction intervals. Several random forest-type algorithms aim at estimating conditional distributions, most prominently quantile regression forests. We propose a novel approach based on a parametric family of distributions characterised by their transformation function. A dedicated novel `transformation tree'' algorithm able to detect distributional changes is developed. Based on these transformation trees, we introduce`transformation forests’‘as an adaptive local likelihood estimator of conditional distribution functions. The resulting models are fully parametric yet very general and allow broad inference procedures, such as the model-based bootstrap, to be applied in a straightforward way. The procedures are implemented in the “trtf” R add-on package currently available from R-forge.

Adaptive Subgroup Selection in Sequential Trials

Tze Leung Lai, Philip W. Lavori, Olivia Liao, Balasubramanian Narasimhan and Ka Wai Tsang

ASSISTant is an R package for a novel group-sequential adaptive trial. The design is motivated by a randomized controlled trial to compare an endovascular procedure with conventional medical treatment for stroke patients; see Lai, Lavori and Liao (Lai 2014). The endovascular procedure may be effective only in a subgroup of patients not known at the design stage but may be learned statistically from the data collected during the course of the trial. The group-sequential design implemented in `ASSISTant` incorporates adaptive choice of the patient subgroup among several possibilities which includes the entire patient population as a choice. Appropriate Type I and type II errors of a test can be defined in this setting and the design maintains a prescribed type I error by using the closed testing principle in multiple testing.

This adaptive design underlies the NIH DIFFUSE-3 trial currently underway. The package is on CRAN and github.

References

Adaptive Choice of Patient Subgroup for Comparing Two Treatments by Tze Leung Lai and Philip W. Lavori and Olivia Yueh-Wen Liao. Contemporary Clinical Trials, Vol. 39, No. 2, pp 191-200 (2014). http://www.sciencedirect.com/science/article/pii/S1551714414001311

useR! 2017 Oral Presentation

The R6 Class System

Winston Chang[1]

1. RStudio

Keywords: Classes, Object-oriented programming, R6, Reference classes

Webpages: https://CRAN.R-project.org/package=R6, https://github.com/wch/R6

R6 is an implementation of a classical object-oriented programming system for *R*. In classical OOP, objects have mutable state and they contain methods to modify and access internal state. This stands in contrast with the *functional* style of object-oriented programming provided by the S3 and S4 class systems, where the objects are (typically) not mutable, and the methods to modify and access their contents are external to the objects themselves.

R6 has some similarities with R's built-in Reference Class system. Although the implementation of **R6** is simpler and lighter weight than that of Reference Classes, it offers some additional features such as private members and robust cross-package inheritance.

In this talk I will discuss when it makes sense to use **R6** as opposed to functional OOP, demonstrate how to use the package, and explore some of the internal design of **R6**.

We R a Community-making high quality courses for high education accessible for all; The >eR-Biostat initiative

Ziv Shkedy[1,2], Nolen Joy Perualila[1], Khangelani Zuma[3], Legesse Debusho[4] and Adetayo Kasim[2,5]

1. *Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, 3590 Diepenbeek, Belgium*
2. *Department of Epidemiology and Biostatistics, Gonder University, Ethiopia*
3. *Human Sciences Research Council (HSRC), PRETORIA, South Africa*
4. *The University of South Africa (UNISA), PRETORIA, South Africa*
5. *Wolfson Research Institute for Health and Wellbeing, Durham University, Durham*

Keywords: Developing countries, master programs, Biostatistucs, E-learning using R

One of the main problems in high education at a master level in developing countries is the lack of high quality course materials for courses in master programs. The **>eR-Biostat** initiative is focused on masters programs in Biostatistics/Statistics and aim to develop new E-learning system for courses at a master level.

The E-learning system, developed as a part of the **>eR-Biostat** initiative, offers free online course materials for master students in biostatistics/statistics in developing countries. For each course, the materials are publicly available and consist of several type of course materials: (1) notes for the course, (2) slides for the course, (3) R programs, ready to use, which contain all data and R code for the all examples and illustrations discussed in the course and (4) homework assignments and exams.

The **>eR-Biostat** initiative introduces a new, R based, learning system, the multi-module learning system, in which the students in the local universities in developing countries will be able to follow courses in different learning format, including e-courses taken online and a combination between e-courses and local lectures given by local staff members. R software and packages are used in all courses as data analysis tool for all examples and illustrations. The **>eR-Biostat** initiative provides a free, accessible and ready to use tool for capacity building in biostatistics/statistics for local universities in developing countries with current low or near zero capacity in these topics. In its nurture, the R community is used for this type of collaboration (for example, CRAN and Bioconductor which offer access to the most up-to-date R packages for data analysis). The **>eR-Biostat** initiative is aimed to bring the R community members for the development of high education courses in the same way it is currently done in software development.

Part II

Lightning Talks

Plasmid Profiler: Comparative Analysis of Plasmid Content in WGS Data

Adrian Zetner[1], Jennifer Cabral[1], Laura Mataseje[1], Natalie Knox[1], Philip Mabon[1], Michael Mulvey[1,2], Gary Van Domselaar[1,2]

1. National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada
2. Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Manitoba, Canada

Keywords: R, Plasmids, Visualization, WGS, AMR,

Webpages: https://cran.r-project.org/web/packages/Plasmidprofiler/index.html, http://biorxiv.org/content/early/2017/03/28/121350.1

Summary: Comparative analysis of bacterial plasmids from short read whole genome sequencing (WGS) data is challenging. This is due to the difficulty in identifying contigs harbouring plasmid sequence data, and further difficulty in assembling such contigs into a full plasmid. As such, few software programs and bioinformatics pipelines exist to perform comprehensive comparative analyses of plasmids within and amongst sequenced isolates. To address this gap, we have developed **Plasmidprofiler**, a pipeline that uses Galaxy and *R* to perform comparative plasmid content analysis without the need for de novo assembly. The pipeline is designed to rapidly identify plasmid sequences by mapping reads to a plasmid reference sequence database. Predicted plasmid sequences are then annotated with their incompatibility group, if known. The pipeline allows users to query plasmids for genes or regions of interest and visualize results as an interactive heat map.

Availability and Implementation: Plasmid Profiler is freely available software released under the Apache 2.0 open source software license.
A stand-alone version of the entire Plasmid Profiler pipeline is available as a Docker container at
https://hub.docker.com/r/phacnml/plasmidprofiler_0_1_6/

The conda recipe for the **Plasmidprofiler** *R* package is available at
https://anaconda.org/bioconda/r-plasmidprofiler

The **Plasmidprofiler** *R* package is also available as a CRAN package at
https://cran.r-project.org/web/packages/Plasmidprofiler/index.html

Galaxy tools associated with the pipeline are available as a Galaxy tool suite at
https://toolshed.g2.bx.psu.edu/repository?repository_id=55e082200d16a504

The source code is available at
https://github.com/phac-nml/plasmidprofiler

The Galaxy implementation is available at
https://github.com/phac-nml/plasmidprofiler-galaxy.

map data from **rnaturalearth** : aiming for sustainability through specialisation and **rOpenSci**

Andy South[1]

1. University of East Anglia and Liverpool School of Tropical Medicine

Keywords: maps, package, sustainability, community

Webpages: https://CRAN.R-project.org/package=rnaturalearth, https://github.com/ropenscilabs/rnaturalearth

rnaturalearth is a new **R** package, accepted to *CRAN* in March this year. It makes Natural Earth map data, a free and open resource, more easily accessible to *R* users. It aims for a simple, reproducible and sustainable workflow from Natural Earth to **rnaturalearth** enabling updating as new versions become available.

rnaturalearth follows from **rworldmap**, a *CRAN* package for mapping world data, which I released more than 7 years ago. **rworldmap** was targetted particularly at relative newcomers to *R*, and has now been downloaded more than 100 thousand times. However, the code is ugly and I haven't had the time to maintain it actively. I have been concerned for a while that making any changes will break it. Now more recently released options such as **tmap** and **choroplethr** are better than **rworldmap** in most respects.

Where **rworldmap** tried to do everything, **rnaturalearth** aims to do fewer things, but to do them better. This approach my be familiar to people. Also being more specialised allows this pacakage to be used in combination with other packages of the users choice.

It is possible to use **rnaturalearth** to have more control over accessing map data, for example specifiying exactly which areas are wanted when dealing with trickiness of countries and dependencies. In this example I use **sp::plot** as a simple, quick way to plot map data, however the output can also be returned as **sf** objects for plotting using other packages.

```
library(rnaturalearth)
library(sp)

# countries, UK undivided
sp::plot(ne_countries(country = 'united kingdom', type = 'countries'))

# map_units, UK divided into England, Scotland, Wales and Northern Ireland
sp::plot(ne_countries(country = 'united kingdom', type = 'map_units'))

# map_units, select by geounit to plot Scotland alone
sp::plot(ne_countries(geounit = 'scotland', type = 'map_units'))

# sovereignty, Falkland Islands included in UK
sp::plot(ne_countries(country = 'united kingdom', type = 'sovereignty'), col = 'red')
```

The package contains pre-downloaded country and state boundaries at different resolutions and facilitates access to other vector and raster data for example of lakes, rivers and roads. Each Natural Earth dataset is characterised on the website according to scale, type and category. **rnaturalearth** will construct the url and download the corresponding file.

```
lakes110 <- ne_download(scale = 110, type = 'lakes', category = 'physical')
sp::plot(lakes110, col = 'blue')
```

I found the early stages of **rworldmap** development a somewhat lonely process. **rnaturalearth** has been through the rOpenSci community open review which improved the code considerably and my experience of developing it. I look forward to this package being more collaborative. I will comment on my experience of issues of community and sustainability within *R* package development.

An R Decision Support Framework for the Identification of BMP in Catchments

Angel Udias[1], *Marco Pastori*[1], *Anna Malago*[1], *Olga Vigiak*[1], *Faycal Bouraoui*[1]

1. *Joint Research Centre, Directorate D - Sustainable Resources Unit D.02 Water and Marine Resource*

Keywords: SWAT, Multi-objective optimization, Best management practices (BMPs), Decision Support System

GREEN-Rgrid is an advanced R version of the model GREEN (Grizzetti, Bouraoui, and Aloe 2012), originally developed in Fortran (Aguzzi, Gasparo, and Macconi 1987) for estimating nutrient loads from diffuse and points sources in Europe. Using R we improved the original structure of the GREEN model ensuring a model fully open to scrutiny since there is a considerable reliance on proprietary software for environmental modelling assessment in Europe (Carslaw and Ropkins 2012). The GREEN-Rgrid model works on a grid cell discretization that the user can change depending on the purpose of the modelling. The model input consists of the latest and best available global data. The GREEN-Rgrid code integrates a landscape routing model to simulate nutrient fluxes of nitrogen-nitrates, total nitrogen, total phosphorous and phosphates across discretized routing units. The grid-based approach was adopted to adapt to the readily available global raster data that can be easily incorporated as model inputs providing a more homogeneous nutrient assessment between different areas of the world. With respect to the original GREEN model, the diffuse source DS was calculated as a function of the gross nutrient balance from agricultural land that is computed as the difference between the inputs (for N: fertilizer application, fixation, and atmospheric deposition; for P: fertilizer application, P release, P transported with sediment) and the output (crop nutrient uptake). A basin attenuation coefficient was applied (aP) to the diffuse sources while two additional coefficients aM and bM are used in the calculation of the in-stream retention (Venohr and et al. 2011). These parameters were calibrated in the model using a Latin Hypercube approach based on packages **FME** and the best parameters set was selected comparing the predicted and observed annual loads using the package **HydroGOF**. Other important packages used for specific purposes in the GREEN-Rgrid model include the package **raster**, **data.table** and **sqldf**. The structure of the model is organized in three folders: the inputs, where the input table and grid raster cell are stored; the Scripts that includes the master R file and the functions and the outputs where all tables and figures generated are saved. We applied the GREEN-Rgrid model in the Mediterranean area (about 8.066x103 km2) using a grid cell size of 5 arc-minute resolution (9.2 km at the equator) and three years of simulations 2005, 2006, 2007. We show the predicted fluxes and concentrations of nutrients in gauged and ungauged grid cells, showing plots and raster maps automatically generated. Finally, some considerations are given for future developments.

useR! 2017 Lightning Talk

Statistics hitting the business front line

Anne Lund Christophersen and Susanna Liberti

Vestas Wind Systems A/S

Keywords: enterprise, collaboration, business, **rmarkdown**.

We will introduce you to a framework we developed to achieve effective collaboration around data analysis in our enterprise environment at Vestas. In this talk we will describe our implementation in *R*, why we chose *R*, which challenges we faced and what we learned during the process.

Setting the scene

We had the task of creating statistical models to be used by the sales teams. Sales already had an *Excel* based tool, and the requirement was that we should continue with this front end. The statistical work would require models developed by a team of people as well as involvement of subject matter experts, hence the framework needed to support collaboration.

On stage

- Sales (end users, 50-100 people around the globe)
- Data analysts and subject matter experts (project team, 10 people in DK + IN)
- In front: Existing *Excel* front end
- In the background: *R*, *GIT*, **rmarkdown**, *SQL*

Scenography

Being in an enterprise world we had to fulfill requirements for maintainability, documentation and reproducibility. At the same time we wanted to achieve i) a code base approach, ii) easier validation methods, iii) automated model deployment and iv) a strong collaborative platform.

Orchestration

On the technical side the main new feature is a self-made package called **harvester**. The **harvester**'s main functions allow us to run markdown-files and fetch selected objects, typically our statistical models.

These fetched models are then wrapped into another internal package called **models** together with interface functions. This is the package used by Sales. The **models** are made available to *Excel* through a self-developed *.NET*-wrapper. In this way the end users will be able to get the most recent models through their normal *Excel* tool.

The collaboration is done through *GIT* where all team members store their analysis *R*-markdowns, shared and validated by subject matter experts. The **harvester** is designed to run the markdowns in *GIT* and fetch the selected output models.

Review

Get inspired on how to integrate validated statistical models into the decision making in the business front line: It is a five star movie.

Reproducible research in computational subsurface hydrology - First steps in R with RMODFLOW and RMT3DMS

Bart Rogiers
Belgian Nuclear Research Centre (SCK • CEN)

Keywords: reproducible research, reproducible reporting, groundwater hydrology, groundwater modelling

Webpages: https://rogiersbart.github.io/RMODFLOW/, https://rogiersbart.github.io/RMT3DMS/

Recently there have been different calls for reproducibility in computational hydrology (*e.g.* Hutton et al. 2016, Fienen and Bakker (2016), Skaggs, Young, and Vrugt (2015)). The use of open-source languages like *R* and *python*, and collaborative coding tools like *Git* may offer a solution (Fienen and Bakker 2016), but only in combination with literate programming the full potential for reproducible research can be reached. With tools like `utils::Sweave` and **knitr**, *R* has been at the forefront of reproducible research in the last few years, and provides a very interesting environment for reproducible research in computational hydrology.

The Environmetrics task view provides a list of different contributed packages relating to surface water hydrology and soil science, but the number of packages dealing with subsurface hydrology remains limited to date. There are packages for creating specific types of plots, like **hydrogeo** which provides Piper diagram (Piper 1944) plotting, or packages for very specific purposes like **quarrint** or **kwb.hantush**.

In order to bring the potential of *R* to computational subsurface hydrology, in the last few years I have been compiling the **RMODFLOW** and **RMT3DMS** packages. These provide interfaces with two of the most-widely used open source codes for groundwater flow and contaminant transport modelling: MODFLOW (Harbaugh 2005) and MT3DMS (Zheng and Wang 1999). Different model input and output file reading functions have been implemented, and different pre- and post-processing tools are available. For visualization of the model data, S3 methods making use of **ggplot2** were implemented as well. The current capabilities of the packages will be demonstrated and examples of reproducible workflows will be provided.

References

Fienen, Michael N., and Mark Bakker. 2016. "HESS opinions: Repeatable research: What hydrologists can learn from the Duke cancer research scandal." *Hydrology and Earth System Sciences* 20 (9): 3739–43. doi:10.5194/hess-20-3739-2016.

Harbaugh, Arlen W. 2005. *MODFLOW-2005, the Us Geological Survey Modular Ground-Water Model: The Ground-Water Flow Process.*

Hutton, C, T Wagener, J Freer, D Han, C Duffy, and B Arheimer. 2016. "Most computational hydrology is not reproducible, so is it really science?" *Water Resources Research* 52: 7548–55. doi:10.1002/2016WR019285.

Piper, Arthur M. 1944. "A Graphic Procedure in the Geochemical Interpretation of Water-Analyses." *Eos, Transactions American Geophysical Union* 25 (6): 914–28. doi:10.1029/TR025i006p00914.

Skaggs, T.H., M.H. Young, and J.A. Vrugt. 2015. "Reproducible Research in Vadose Zone Sciences." *Vadose Zone Journal* 14 (10): 0. doi:10.2136/vzj2015.06.0088.

Zheng, Chunmiao, and Patrick Wang. 1999. "MT3DMS, A modular three-dimensional multi-species transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems; documentation and user's guide." *U.S. Army Engineer Research and Development Center Contract Report SERDP-99-1, Vicksburg, MS*, 202+.

Using R for optimal beer recipe selection

Benjamin Høyer[1,2,3]

1. Think Big Analytics, a Terdata Company
2. Dynamo Brews
3. White Labs

Keywords: Brewing, beer, yeast, visualisation, summary statistics

Webpages: https://dynamobrew-stats.shinyapps.io/WhiteLabsBrewAppHb/

How is *R* helping brewers to choose the best yeast for their beer? How does yeast choice influence predicted quantities like bitterness versus measured bitterness?

This is meant as a short, fun presentation, touching on Beglian brewing heritage.

I will present walk-through of an interactive **shiny** app I created for a client in the brewing industry.

The initial task was primarily to ingest the data and produce an interactive enviroment, where the client's employees could explore their data. I will not spend too much time on this, but will mention briefly the database technologies used to access the data. I will mention some of my experiences with productionalised a full data flow (ingestion, transformations, outlier handling, visualisation).

The main goal of the presentation will be to visually demonstrate differences between beer styles with regard to their recipes, and to demonstrate the importance of matching beer style with an appropriate yeast.

Before the Conference, I plan on implementing a clustering method based on Self-Organising Maps. This should be a very nice way to explore the natural clustering of recipes – and it should map very neatly to beer styles.

Pending approval from RateBeer, I will also join some high level beer styles data scraped from the https://www.ratebeer.com website over the brewing yeast data.

BivRegBLS, a new *R* package: Tolerance Intervals and Errors-in-Variables Regressions in Method Comparison Studies

Bernard G FRANCQ[1,2] *and Marion BERGER*[3]

1. *UCL, ISBA, Belgium*
2. *GSK, Rixensart, Belgium*
3. *Sanofi, Montpellier, France*

Keywords: Tolerance Intervals, Method Comparison Studies, Agreement, Errors-in-Variables Regression, Bivariate Least Square

Webpages: https://CRAN.R-project.org/package=BivRegBLS

The need of laboratories to quickly assess the quality of samples leads to the development of new measurement methods. These methods should lead to results comparable with those obtained by a standard method.

Two main methodologies are presented in the literature. The first one is the Bland-Altman approach with its agreement intervals (AIs) in a (M=(X+Y)/2,D=Y-X) space, where two methods (X and Y) are interchangeable if their differences are not clinically significant. The second approach is based on errors-in-variables regression in a classical (X,Y) plot, whereby two methods are considered equivalent when providing similar measures notwithstanding the random measurement errors. These methodologies can be used in many other domains than clinical.

During this talk, novel tolerance intervals (TIs) (based on unreplicated or replicated designs) will be shown to be better than AIs as TIs are easier to calculate, easier to interpret, and are robust to outliers. Furthermore, it has been shown recently that the errors are correlated in the Bland-Altman plot. The coverage probabilities collapse drastically and the biases soar when this correlation is ignored. A novel consistent regression, CBLS (Correlated Bivariate Least Square), is then introduced. Novel predictive intervals in the (X,Y) plot and in the (M,D) plot are also presented with excellent coverage probabilities.

Guidelines for practitioners will be discussed and illustrated with the new and promising *R* package **BivRegBLS**. It will be explained how to model and plot the data in the (X,Y) space with the BLS regression (Bivariate Least Square) or in the (M,D) space with the CBLS regression by using **BivRegBLS**. The main functions will be explored with an emphasis on the output and how to plot the results.

References

BG Francq, B Govaerts (2016). How to regress and predict in a Bland-Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. Statistics in Medicine, 35:2328-2358.

rdwd - manage German weather observations

Berry Boessenkool[1]

1. Potsdam University, Germany

Keywords: Weather, Data, Climate, Germany, FTP

Webpages: https://CRAN.R-project.org/package=rdwd, https://github.com/brry/rdwd#rdwd

Since 2014, the German weather service (Deutscher Wetterdienst, DWD) has released over 25'000 observational weather records from stations across Germany. Along with several derived and gridded datasets, they are available free of charge on the Climate Data Center FTP server.

This vast amount of data can be hard to systematically search and inspect. The *R* package **rdwd** provides infrastructure to select and download/process data. For **data selection**, a list of stations and files is provided along with an interactive map and a convenient query function called `selectDWD`. Vectorized **data downloading and reading** is made easy through `dataDWD` and `readDWD`, including the option to resume aborted downloads, correctly read the (differently structured) fixed width column files and convert time stamps to time/date columns.

The purpose of **rdwd** is to facilitate usage and analysis of weather data in Germany. The targeted user group contains many scientists who may not be very familiar with *R*. Therefore an emphasis was put on clear vignettes and an instructive github readme file. **rdwd** was first released in January 2017 and with some additional features, it is now ready for broader advertisement.

Graduate from plot to ggplot2: Using R to visualize the story of Ebola survivors in the PREVAIL III Ebola Natural History Study

Bionca M Davis[1] *and Sienneh Z Tamba*[2]

1. *University of Minnesota, Twin Cities, MN*
2. *Partnership for Research on Ebola Virus in Liberia (PREVAIL), Monrovia, Liberia*

Keywords: R, ggplot2, visualization, research, observational study

Webpages: https://CRAN.R-project.org/package=ggplot2

In late 2013, Ebola Virus Disease began in Guinea; it later spread to Liberia and subsequently to Sierra Leone in early 2014. Although the three countries were declared Ebola-Free almost a year ago, Ebola survivors are still struggling with lingering issues. In order to better understand post-Ebola sequelae, a Liberian-US partnership called the Partnership for Research on Ebola Virus in Liberia (PREVAIL) is in the midst of a large observational study that plans to follow Ebola survivors and their close contacts for up to 5 years. In this lightning talk, we will use R's visualization capabilities to guide you through the struggle of Ebola survivors as told by the data from the PREVAIL study. No crowded tables with a long list of symptoms and p-values, but rather beautiful visualizations created using the gglot2 package. At the end of this talk, it is hoped that you will be able to answer questions such as: Where in Liberia do survivors come from? What were their professions before the epidemic? How sick were they? Have they reached complete physical, social, and mental well-being? In the process, we hope you will be compelled to use ggplot2 to increase the overall quality of data visualizations in your reports.

nsoAPI - retrieving data from National Statistical Offices with R

Bo Werth[1]

1. OECD

Keywords: API, CURL, Official Statistics

Webpages: http://rdata.work/slides/nsoapi/

National Statistical Offices have started setting up web services to provide published information through data APIs. Even though international standards exist, e.g. SDMX, the majority of NSOs create their individual API and few use existing community standards.

nsoAPI is an attempt to create a single package with functions for each provider that convert a custom data format into an R standard time series format ready for analysis or further transformation.

("Opendata Tables" 2015) lists tables that can be retrieved from SDMX (International Organizations, ABS: Australia, INEGI: Mexico, INSEE: France and ISTAT: Italy, NBB: Belgium), the pxweb package (PXNET2: Finland, SCB: Sweden) and the nsoAPI package (BEA: USA, CBS: the Netherlands, GENESIS: Germany, ONS: UK, SSB: Norway, STATAT: Austria, STATBANK: Denmark).

With the exception of France, large countries tend to set their own standards. The BEA (USA) and ONS (UK) require the user to create an ID that needs to be submitted with each request. GENESIS (Germany) require the user to pay 500 Euros per year (250 Euros for academic users) to access the API.

References

"NsoApi Vignette." 2015. https://github.com/bowerth/nsoApi/blob/master/vignettes/nsoApi.md.

"Opendata Tables." 2015. https://www.gitbook.com/read/book/bowerth/opendata-tables.

OpenSpat, spread the spatial world

Brostaux Yves[1] and Fontez Bénédicte[2]

1. Gembloux Agro-Bio Tech, University of Liege
2. MISTEA, Montpellier SupAgro

Keywords: spatial data, lifelong learning, open source

Webpages: http://www.openspat.eu

The availability of spatial data is increasing every year, with the growing access to free satellite imagery, to cheap drone cameras and the omnipresence of GPS on all sort of devices, from our phones to farm tractors. Many applications in agriculture and environmental sciences, which are often dealing with the organization of space and territories, can profit from those new sources of data, like precision farming, land-use use planning, environmental monitoring,etc.

Those information can be extracted, analysed and turned into models to support better decisions: apply the right amount of fertilizer at the right place in the field, predict the future extension of an urban zone, draw a map of the flooding or fire risks. But to do that, people need skills and tools to access, extract, explore and analyse such data.

That's why we started a project to design a lifelong learning course on spatial data analysis at the european level, with three partners (University of Liege, University of Lisboa and Montpellier SupAgro), based on free and open tools like *QGIS* and *R* so that everybody can install and use them, because knowledge access shouldn't be restricted to organizations which can afford costly licenses.

We will present the construction of the program and the pedagogical approach we'll use. Blended learning will be tested for a european master level course for adult learning.

eventstudies: An *R* package for conducting event studies and a platform for methodological research on event studies

Chirag Anand,[1] Vikram Bahure,[2] Vimal Balasubramaniam,[3] and Ajay Shah[1]

1. National Institute of Public Finance and Policy, India
2. Bocconi University, Italy
3. PhD Candidate, University of Oxford

Keywords: event study methodology

Webpages: https://CRAN.R-project.org/package=eventstudies, https://github.com/nipfpmf/eventstudies

This *R* package allows a dataset to be studied in an event-time frame and perform parametric/non-parametric analysis using several inference procedures. There are currently three adjustment functions and three inference strategies including the classical event study using market model and t-test (Brown and Warner 1985) along with an implementation of Augmented Market Model (Patnaik and Shah 2010). The package contains a user-friendly all encompassing function `eventstudy` to conduct an event study in one line of *R* code with other functions to provide more flexibility and control. It can also be used to develop novel research methods in event studies (Patnaik, Shah, and Singh 2013).

References

Brown, Stephen J, and Jerold B Warner. 1985. "Using Daily Stock Returns: The Case of Event Studies." *Journal of Financial Economics* 14 (1). Elsevier: 3–31.

Patnaik, Ila, and Ajay Shah. 2010. "Does the Currency Regime Shape Unhedged Currency Exposure?" *Journal of International Money and Finance* 29 (5): 760–69.

Patnaik, Ila, Ajay Shah, and Nirvikar Singh. 2013. "Foreign Investors Under Stress: Evidence from India." *International Finance* 16 (2): 213–44.

The cutpointr package: Improved and tidy estimation of optimal cutpoints

Christian Thiele[1] *and Gerrit Hirschfeld*[1]

1. Osnabrück University of Applied Sciences

Keywords: cutpoints, tidy

Webpages: https://github.com/Thie1e/cutpointr

Clinicians often use cutpoints or decision-thresholds to decide e.g. whether or not a patient with a depression score of "20" needs treatment for her or his depression. The *R* package **cutpointr** allows for estimating such optimal cutpoints for binary decisions by maximizing a specified metric or by using kernel estimation or distribution-based methods (Fluss, Faraggi, and Reiser 2005; Leeflang et al. 2008). The package includes a parallelizable bootstrapping routine to provide estimates of the cutpoints' variability and their out-of-bag performance. **cutpointr** follows current tidy programming practices to allow for efficient estimation and use in simulation studies as well as interplay with functions from the tidyverse. Furthermore, **cutpointr** accepts user defined functions and includes existing functions (López-Ratón et al. 2014) to calculate optimal cutpoints. We also discuss future plans for **cutpointr**, specifically a Shiny-interface to make the package more accessible.

References

Fluss, Ronen, David Faraggi, and Benjamin Reiser. 2005. "Estimation of the Youden Index and Its Associated Cutoff Point." *Biometrical Journal* 47 (4): 458–72. http://onlinelibrary.wiley.com/doi/10.1002/bimj.200410135/full.

Leeflang, Mariska MG, Karel GM Moons, Johannes B. Reitsma, and Aielko H. Zwinderman. 2008. "Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions." *Clinical Chemistry*, no. 4: 729–38. http://go.galegroup.com/ps/i.do?id=GALE%7CA209106300&sid=googleScholar&v=2.1&it=r&linkaccess=fulltext&issn=00099147&p=AONE&sw=w.

López-Ratón, Mónica, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. 2014. "OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests." *Journal of Statistical Software* 061 (i08). http://econpapers.repec.org/article/jssjstsof/v_3a061_3ai08.htm.

R in Minecraft

David Smith[1]

1. Microsoft

Keywords: R, Minecraft, education

Minecraft is an open-world creativity game, and a hit with kids. To get kids interested in learning to program with R, a team at the ROpenSci unconference created the “miner” package. Kids can use this package interact with the Minecraft world with simple R commands, and learn to use R with the help of a companion book.

Plot Colour Helper - Finally an easy way to pick colours for your R plots!

Dean Attali

Keywords: ggplot2, visualization, shiny, colours, plotting

Webpages: https://github.com/daattali/colourpicker, https://cran.r-project.org/package=colourpicker

You've just made an amazing plot in *R*, and the only thing remaining is finding the right colours to use. Arghhh this part is never fun… Yu're probably familiar with this loop: try some colour values -> plot -> try different colours -> plot -> repeat. Don't you wish there was a better way?

Well, now there is :)

If you've ever had to spend a long time perfecting the colour scheme of a plot, you might find the new Plot Colour Helper handy. It's an RStudio addin that lets you interactively choose combinations of colours for your plot, while updating your plot in real-time so you can see the colour changes immediately.

Data Error! But where?

Edwin de Jonge[1] and Mark van der Loo[1]

Keywords: Error Datacleaning

Webpages: https://CRAN.R-project.org/package=errorlocate, https://github.com/data-cleaning

An important but undermentioned activity needed for statistical analysis is data-cleaning. No measurement is perfect, so data often contain errors. Obvious errors e.g. negative age are easily detected, but observations that contain variables that are logically related e.g. marital status and age are more tricky. R package `errorlocate` allows for pin pointing errors in observations using the Feligi-Holt algorithm and validation rules from R package `validate`. The errors can automatically be removed using a pipe-line syntax.

useR! 2017 Lightning Talk

Preparing Datetime Data with Padr

Edwin Thoen

Data Scientist Rabobank

Keywords: Data Preparation, Datetime Data, Tidyverse,

Webpages: https://CRAN.R-project.org/package=padr, https://github.com/edwinth/padr

`padr` solves two problems that you can be confronted with when preparing datetime data for analysis. First, data is often recorded on too fine a granualarity. For instance, the timestamp registers the moment up to the second, while you want to do the analysis on an hourly level. The `thicken` function will add a column to a data frame. In conjunction with `dplyr`, it will allow for quick aggregation to the higher level. Second, when no events take place there are typically no data records generated. This is sensible from a storage perspective, but often unhelpful for analysing the data. In this context the `pad` function is used. Besides demonstrating these two functions, I will elaborate on the concept of the *interval*, on which both functions heavily rely.

useR! 2017 Lightning Talk

Automatic Machine Learning in *R*

Erin LeDell[1]

1. H2O.ai

Keywords: Machine Learning, Ensemble Learning, Automatic Machine Learning, Black-Box Optimization, Distributed Computing

Webpages: https://CRAN.R-project.org/package=h2o, https://gitbub.com/h2oai/h2o-3

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software that can be used by non-experts. The first steps toward simplifying machine learning in *R* focused on developing simple, unified interfaces to a variety of machine learning algorithms. This effort also involved providing a robust toolkit of utility functions that perform common tasks in machine learning such as random data partitioning, cross-validation and model evaluation. Successful examples of this simplification effort include the **caret**, **mlr** and **h2o** *R* packages.

Although these tools have made it easier for non-experts to experiment with machine learning, there is still a fair bit of knowledge and background in data science that is required to produce high-performing, production-ready/research-grade machine learning models. Deep Neural Networks in particular (which have become wildly popular in the past five years) are notoriously difficult for a non-expert to tune properly. In order for machine learning software to truly be accessible to non-experts, such systems must be able to automatically perform proper data pre-processing steps and return a highly optimized machine learning model.

H2O.ai has developed a distributed Automatic Machine Learning system called H2O AutoML (to be officially released in the **h2o** *R* package (H2O.ai 2017) approx. May-June 2017; currently in pre-release), which will be the first open source Automatic Machine Learning system available in *R*. In this presentation, we will present our methodology for automating the machine learning workflow, which includes feature pre-processing and automatic training and tuning of many models within a user-specified time-limit. The user can also specify which model performance metric that they'd like to optimize and use a metric-based stopping criterion for the AutoML process rather than a specific time constraint. By default, stacked ensembles will automatically trained on subset of the individual models to produce a highly predictive ensemble model, although this can be turned off if the user prefers to return singleton models only.

The interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time-constraint. Below is an example of how to specify an AutoML run for the default run-time.

```
aml <- h2o.automl(training_frame = train, response_column = "class")
```

The AutoML object includes a history of all the data-processing and modeling steps that were taken, and will return a "leaderboard" of all the models that were trained in the process, ranked by a user's model performance metric of choice.

References

H2O.ai. 2017. *H2O R Package*. https://github.com/h2oai/h2o-3/tree/master/h2o-r.

useR! 2017 Lightning Talk

Functional Input Validation with valaddin

Eugene Ha[1]

1. Feingold Technologies GmbH

Keywords: input validation, type safety, defensive programming, functional programming

Webpages: https://CRAN.R-project.org/package=valaddin, https://github.com/egnha/valaddin

To cope with the everyday hazards of invalid function inputs, *R* provides the functions `stop()` and `stopifnot()`, which can express input requirements as show-stopping assertions. While this way of validating inputs is both straightforward and effective, its rigidity as a fixture of a function, and its tendency to clutter code, add inertia to the process of interacting and programming with data.

In this talk, we demonstrate a more nimble take on input validation using the valaddin package, which address these shortcomings by viewing input validation as a *functional* transformation. We explore concrete use cases to illustrate the flexibility and benefits of this alternative approach.

ROI - R Optimization Infrastructure

Florian Schwendinger[1] and Stefan Theussl[2] and Kurt Hornik[1]

1. WU Vienna University of Economics and Business
2. Raiffeisen RESEARCH

Keywords: optimization, mathematical programming

Webpages: https://cran.r-project.org/web/packages/ROI/index.html, https://r-forge.r-project.org/projects/roi/

Optimization plays an increasingly important role in statistical computing. Typical applications include, among others, various types of regression, classification and low rank matrix approximations. Due to its wide application there exist many resources concerned with optimization. These resources involve software for modeling, solving and randomly generating optimization problems, as well as optimization problem collections used to benchmark optimization solvers. The *R* Optimization Infrastructure package **ROI** bundles many of the available resources used in optimization into a unified framework. It constitutes a unified way to formulate and store optimization problems by utilizing the rich language features *R* has to offer, rather than creating a new language. In **ROI** an optimization problem is stored as a single object, which ensures that it can be easily be saved and exchanged. Furthermore, the streamlined construction of optimization problems combined with a sophisticated plugin structure allows package authors and users to exploit different solver options by just changing the solver name. Currently, the **ROI** plugins include solvers for general purpose nonlinear optimization as well as for linear, quadratic and conic programming. Additionally, plugins for reading and writing optimization problems in various formats (e.g. `MPS`, `LP`) and plugins for problem collections (e.g. `netlib`, `miplib`) transformed into the **ROI** format are available.

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

F. Privé[1], H. Aschard[2] and M.G.B. Blum[1]

1. Université Grenoble Alpes, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG, UMR 5525, Grenoble, France
2. Département de Génomes et Génétique, Institut Pasteur, Paris, France

Keywords: Statistics, Big Data, Memory-mapping, Parallelism

Webpages: https://github.com/privefl/bigstatsr

The *R* package **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.

useR! 2017 Lightning Talk

R and Tableau Integration: A Case approach

Francis Nguyen[1], *Sean Lam*[1], *Marcus Ong*[2]

1. Health Services Research Center, SingHealth, Singapore
2. Department of Emergency Medicine, Singapore General Hospital, Singapore

Introduction

R is a powerful statistical engine comes with extensive libraries and innovative methodology implementation. On the other hand, Tableau is a popular business intelligence tool to visualize graphs and chart in fairly easy and straight forward manner. Tableau and R complement each other in areas where heavy data crunching is needed for visualization or when geospatial visualization and geodata manipulation are both needed. In this talk, two cases will be discussed to highlight the complementary merits of both software: dot map and choropleth map.

Mapping

Geospatial community is familiar with a laundry list of libraries and packages to be attached before perform spatial statistics. However, the users' contribution to the CRAN mirrors has made monumental advances in bridging academic area and industry, shorten the gap between research and implementation. Both essential geographic data management and analysis can be performed readily in R. These advantages empower R as a leading open-source programming language against expensive commercial software in the market.

Despite extensive libraries and state-of-the-art algorithm available in R, visualization is not always praised as the language's forte. For example, the omnipotent plot command to display kernel density estimation (KDE) in R challenges analysts to be well conversant with the location details to answer the key question: where is the hotspot? User interactivity functionality is another hurdle to overcome. The reactive graph requires considerably substantial amount of codes and sometimes cumbersome JavaScript to carry out simple customization or overlay dots on the base map. Some may argue that the communicating the point pattern process result with R is not an easy task. It is simply not compelling and visually aesthetic enough. #Integrating R and Tableau to produce map for spatial statistics Even though Tableau is hardly mentioned when it comes to mapping and geospatial, the user friendliness and interactivity are its selling points. Presenter has found majority of his mapping requirements satisfied by Tableau's functionalities. The talk will detail two applications where R and Tableau can come together and greatly complement each other.

- Display a set of point objects on a map: in this example, a collection of coordinates will be projected to the standard WGS 1984 Coordinate Reference System and Tableau acts as a base map to overlay the points.
- Juxtapose and highlight region with using areal data: the task requires analyst to map area objects, compare different regions and highlight member of top and bottom groups according a ranking parameter. Technically speaking, a choropleth map is needed.

Gamifyr: Transforming Machine Learning Tasks into Games with Shiny

Giorgio Maria Di Nunzio[1]
1. Department of Information Engineering

Keywords: Gamification, Text Classification, R Shiny, Interactive Machine Learning.

Webpages: https://gmdn.shinyapps.io/Classification/

Supervised machine learning algorithms require a set of labelled examples to be trained; however, the labelling process is a costly, time consuming task. In the last years, mixed approaches that use crowd-sourcing and interactive machine learning (Amershi et al. 2014) have shown that it is possible to create annotated datasets at affordable costs (Morschheuser, Hamari, and Koivisto 2016). One major challenge is to design a system that motivates people to participate in these labelling tasks.

In this context, 'gamification' has become popular, i.e. 'the use of game design elements in non-game contexts' (Deterding et al. 2011). For instance, an increasingly common feature of online communities and social media sites is a mechanism for rewarding user achievements based on a system of badges and points. They have been employed in many domains, including educational sites like Khan Academy, and tourist review sites like Tripadvisor. At the most basic level, these game elements serve as a summary of a user's key accomplishments; however, experience with these sites also shows that users will put in non-trivial amounts of work to achieve particular badges, and as such, badges can act as powerful incentives (Anderson et al. 2013).

In this work, we present recent studies of gamification in text classification and the development of a Shiny application (Di Nunzio, Maistro, and Zilio 2016). This application, initially designed to understand probabilistic models, was redesigned as a game to gather labelled data from lay people during the European Researchers' Night in September 2016 at the University of Padua. We have tested this application with two goals in mind: i) how gamification can be used to understand what is the 'price' of labelling a small amount of objects for building a reasonably accurate classifier, ii) to analyze the classification performance given the presence of small sample sizes and little training. We will describe three different interfaces and the analysis of the results: a pilot experiment with PhD and post-doc students, a second experiment with primary and secondary school students, and a third experiment with a computer installed in a bank.

References

Amershi, Saleema, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. "Power to the People: The Role of Humans in Interactive Machine Learning." *AI Magazine* 35 (4): 105–20. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2513.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. "Steering User Behavior with Badges." In *Proceedings of the 22Nd International Conference on World Wide Web*, 95–106. WWW '13. New York, NY, USA: ACM. doi:10.1145/2488388.2488398.

Deterding, Sebastian, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. "From Game Design Elements to Gamefulness: Defining 'Gamification'." In *Proc. of the 15th International Academic Mindtrek Conference: Envisioning Future Media Environments*, 9–15. MindTrek '11. New York, NY, USA: ACM. doi:10.1145/2181037.2181040.

Di Nunzio, Giorgio Maria, Maria Maistro, and Daniel Zilio. 2016. "Gamification for Machine Learning: The Classification Game." In *Proceedings of the Third International Workshop on Gamification for Information Retrieval Co-Located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.*, 45–52. http://ceur-ws.org/Vol-1642/paper7.pdf.

Morschheuser, B., J. Hamari, and J. Koivisto. 2016. "Gamification in Crowdsourcing: A Review." In *2016 49th Hawaii International Conference on System Sciences (Hicss)*, 4375–84. doi:10.1109/HICSS.2016.543.

R in a small-sized bank's risk management

Goran Lovric ^{1} and Martin Hoelbl ^{1},

1. Western Union Internation Bank GmbH

Keywords: risk management, shiny, automation, risk modelling, reporting

The typical risk management in a small-sized bank is heavily dependent on manual processes and using well-known spreadsheet applications even far beyond their original scope, like being used as a data storage tool. This has several reasons: First, spreadsheet applications (ie. Microsoft Excel) are well known and distributed throughout different countries and industries. Further risk management often receives data and information from all kind of different departments and hence has to deal with diverse data systems and data structures. Nevertheless, most of those systems are able to export their information to an Excel-compatible format. Finally, the costs to invest in systems and tools that would adequately replace classical spreadsheet applications are usually too high for small-sized banks.

Nevertheless, there are some problems attached to this procedure. This talk will show those weaknesses while proposing a different solution which allows to automate as much as possible and to integrate very important features to the risk management. This new approach which uses different functionalities of *R* is focused on three major topics:

- Automation of reporting processes - increasing efficiency
- Risk modelling - one step ahead of spreadsheet applications
- Present and distribute the results - Impress your stakeholders

The first part focuses on different practical examples that were done by the reporting department in Excel (e.g. Margin Call reports, exposure calculations,etc.). Those processes were done mostly manually and are therefore very error-prone. An elegant solution is using the different functionalities in *R* to connect to different systems (Salesforce, Oracle,.) and the possibility to run *R* scripts as a batch in the task scheduler and hence free the time of the analysts to focus on the qualitative part of the topics. Further the usage of **shiny** allows to standardize processes that need user inputs and helps to improve the user experience.

The second topic shows the usage of BI methods using similar methods as in the first part (data migration/ connecting systems). This includes the development of a rating model for SME and Corporate clients and a Value-at-risk model for FX derivatives as an example. Both cases are fully executable *R* codes which provide full audit trail.

Part three focuses on the very important topic of presenting 'results' (ie. reports, models, etc.) to different stakeholders. Here some newer developments in the *R* community like **shiny** and **shinydashoard** show its full potential. In addition to **shiny** being used for rating tools it can be used to show different type of reports without 'frightening' other stakeholders with the typical *R* environment.

To sum up there is a wide field of applications for using *R* in the risk management which improves the overall performance. The talk will be accompanied by a live demo of the tools discussed.

Effectively Remember R-Skills with Spaced Repetition

(An Introduction to the r2anki-package)

Henning Bumann and Malte Grosser[1]

1. University Medical Center Hamburg-Eppendorf

Keywords: Spaced Repetition Learning, r2anki, RMarkdown

Webpages: https://github.com/henningsway/r2anki

When you learn and use *R* you need to memorize important commands to solve programming tasks effectively. Unfortunately some less frequently used function calls can be forgot quite easily as you learn more about the language.

Spaced repetition learning offers a solution to this problem by exposing you only to learning content, that you are about to forget. The open source software *Anki* offers a fantastic cross-plattform implementation of this approach.

The lighning talk will briefly introduce the idea of spaced repetition learning and how the `r2anki`-package can be used to easily convert RMarkdown-scripts into a set of Anki-flashcards, that can be shared among the commmunity.

References

- https://en.wikipedia.org/wiki/Spaced_repetition
- https://ankisrs.net

useR! 2017 Lightning Talk

bsplus: Using Twitter Bootstrap to extend your Shiny app

Ian Lyttle^1

1. Analytics Applications and Programs

Keywords: shiny, rmarkdown, HTML, Twitter Bootstrap

Webpages: https://cran.r-project.org/package=bsplus, http://ijlyttle.github.io/bsplus/

With the advent of **shiny** (Chang et al. 2017) modules, you can create and support apps with more components and more complexity. One of the limiting factors is that we have but one "dimension" of interfaces using a tabsetPanel in the UI. This was the motivation to develop a second, independent "dimension" of interfaces in an "accordion-sidebar" framework. This is one of the function families provided in the **bsplus** (Lyttle 2017) package.

As well, the **bsplus** package lets you compose HTML using pipes. Its functions are designed to help you access Twitter Bootstrap (Bootstrap Core Team 2017) components independent of the server side of **shiny**. It also includes collapsible panels, accordions, carousels, tooltips, popovers and modals. You can use carousels to contain and display images (plots), whereas tooltips, popovers and modals can be useful for providing help and documentation for your apps.

References

Bootstrap Core Team. 2017. *Bootstrap: The World's Most Popular Mobile-First and Responsive Front-End Framework*. https://getbootstrap.com.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2017. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Lyttle, Ian. 2017. *Bsplus: Adds Functionality to the R Markdown + Shiny Bootstrap Framework*. https://CRAN.R-project.org/package=bsplus.

simmer: Discrete-Event Simulation for R

Iñaki Úcar[1] and Bart Smeets[2]

1. Universidad Carlos III de Madrid
2. dataroots

Keywords: Discrete-Event Simulation

Webpages: https://CRAN.R-project.org/package=simmer, http://r-simmer.org/

Discrete-Event Simulation (DES) is a powerful modelling technique that breaks down complex systems into ordered sequences of well-defined events. Its applications are broad (from process design, planification and optimisation to decision making) in a wide range of fields, such as manufacturing, logistics, healthcare and networking.

This talk presents **simmer**, a package that brings DES to *R*. It is designed as a generic yet powerful process-oriented framework. The architecture encloses a robust and fast simulation core written in *C++* with integrated monitoring capabilities, allowing for easy access to time series data on processes and resources. It provides a rich and flexible *R* API that revolves around the concept of a trajectory, a common path in the simulation model for entities of the same type. A trajectory can be defined as a recipe-like set of activities that correspond to common functional DES blocks. These activities are exposed as intuitive verbs (e.g., `seize`, `release` and `timeout`) and chained using the popular pipeline notation `%>%`, which makes for clear and transparent DES modelling.

Over time, the **simmer** package has seen significant improvements and has been at the forefront of DES for *R*.

Working with R when internet is not reliable

Jens Carl Streibig
University of Copenhagen,
Plant and Environmental Sciences

Keywords: miniCRAN, Agricultural Sciences, Statistics

Webpages: https://CRAN.R-project.org/package=miniCRAN, http://rstats4ag.org/

R has evolved over time and currently consists of more than 10,000 packages. Virtually any aspects of statistical methods e.g., in the agricultural sciences, are readily available. Packages can be accessed anywhere, also in countries with few resources for purchasing commercial programmes (e.g., SAS, SPSS, Matlab). However, when in regions where seamless running internet is an exception rather than the rule we have problems. Introducing *R* and relevant packages to new students should not begin with struggling downloading the packages. In training and instruction situations, the package **miniCRAN** can help us maintain a private mirror of a subset of packages that are relevant to distribute among students irrespective of the functioning of the internet. In addition, the package **miniCRAN** makes it possible to make a dependency tree for a given set of packages. An important facility is the capability to download older version packages from the CRAN archives. However, dependencies for old package versions cannot be determined automatically and the end user must specify.

R.gov: making R work for government

Jeremy Darot
Scottish Government

Keywords: Government, Public Sector, Official Statistics

Webpages: https://rdotgov.wordpress.com/

Over the last decade, government organisations around the world have increasingly adopted *R* for their analytical needs, driven by the promise of more powerful and reproducible data analysis pipelines, Shiny - and lower costs. *R* maturity varies considerably across the public sector, with some organisations just starting to experiment with *R*, and others already using it as their primary workhorse for official statistics production and dissemination (Templ and Todorov 2016).

We will outline the main barriers to introducing *R* in government organisations, from IT culture to the career progression of statisticians, and how the Scottish Government is overcoming them.

We will also present R.gov, an informal group open to all public sector organisations which aims to enable and promote the use of *R* in government. The group already has members in ten countries, and provides a forum for sharing knowledge and fostering collaborations.

We hope this lightning talk will spark productive conversations and help create new connections, not just within government, but across the entire *R* community.

References

Templ, Matthias, and Valentin Todorov. 2016. “The Software Environment R for Official Statistics and Survey Methodology.” *Austrian Journal of Statistics* 45: 97–124. doi:10.17713/ajs.v45i1.100.

R Blogging with blogdown and GitHub

Joseph Rickert[1], Steven Fazzio[1], and Amanda Gadrow[1]

1. RStudio

Keywords: R blogging, blogdown, GitHub, R Markdown

Webpages: https://www.rstudio.com/rviews/, https://github.com/rstudio/blogdown

Blogging about *R* presents its own technical challenges. The need to include sophisticated *R* code, Shiny applications, R Markdown documents, and interactive graphics severely taxes traditional blogging platforms such as WordPress and Typepad. We believe that the new **blogdown** package, which generates static websites using R Markdown and Hugo, represents the future of *R* blogging. In this talk, we will describe the basics of the **blogdown** package, and share our experiences editing and producing RStudio's R Views blog using **blogdown** as the blog engine and GitHub as the platform for coordination.

Application of R and Shiny in multiomics understanding of blood cancer biology and drug response

Junyan Lu[1], *Małgorzata Oleś*[1], *Sascha Dietrich*[2],[3], *Sina Oppermann*[2], *Sebastian Scheinost*[2], *Thorsten Zenz*[2] *and Wolfgang Huber*[1]

1. *European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
2. *Molecular Therapy in Hematology and Oncology, and Department of Translational Oncology, National Center for Tumor Diseases and German Cancer Research Centre, Heidelberg*
3. *Department of Medicine V, University Hospital Heidelberg, Heidelberg, Germany*

Keywords: multiomics, drug screen, blood cancer, personalized medicine, shiny

Webpages: https://github.com/lujunyan1118/DrugScreenExplorer

Better tools for response prediction would improve quality of cancer care. To gain further insight into the pathogenesis of blood cancers as well as to understand determinants of drug response, we measured the sensitivity of primary tumor samples from a large cohort of leukemia/lymphoma patients to marketed drugs and chemical probes. Alongside, genome, transcriptome, DNA methylome and metablome data were obtained for the same set of patient samples, providing a valuable multidimensional resource for blood cancer study.

To facilitate the query and analysis of our dataset, we have created an R and Shiny based online platform – DrugScreenExplorer. This platform incorporates various tools for quality assessment, data visualization, exploratory data analysis and association test. For example, the drug screening quality can be readily examined by interactive heatmap plots of the screening plates and outlier samples and drugs can be detected by unsupervised clustering methods, such as principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). Moreover, associations among different omics datasets can be analyzed and visualized within this platform, facilitating hypothesis generation and subsequent experimental validation.

Those handy tools enable us to achieve seamless and efficient collaboration between dry lab and wet lab groups and to extract useful information from out multi-layer structure dataset in order to gain insight into the complexity of drug response and genotype-phenotype relationships in cancer. Currently, this Shiny platform are customized for our in-house data. But with further extensions, such as allowing users to upload their own data, it can be used as general-purpose tools to streamline the pre-processing, quality control, data visualization and reporting for other drug screening projects as well.

useR! 2017 Lightning Talk

Use of templates within an R package to create a (semi-)automated analysis workflow and/or report

Kirsten van Hoorde[1] and Laure Cougnaud[1]

1. OpenAnalytics

Keywords: reporting, reproducibility, automation, Rmarkdown

Have you ever had the feeling that the creation of your data analysis report(s) resulted in quite some copy-paste from previous analyses? This copy-pasting is time-consuming and prone to errors.

If you need to analyze frequently quite similar data, e.g. from a standardized workflow or different experiments on a specific platform, automation of your analysis can be a useful and time-saving step.

An efficient solution might be the development of modular template documents integrated in an R 'template' package. This package contains the common analysis parts consistent throughout the different analyses, in different child (potentially nested) template documents (module). These templates can be seen as the equivalent of an R function, integrated within an R package, for reporting with input parameters and potentially some default values (necessary/specific analysis parts).

A main 'master' template, specific for each analysis (e.g. experiment) can call the child document(s) contained in the package. It is advisable and user-friendly, for yourself and potential other users, to create a start template for this master document, where the required and optional input parameters necessary for the downstream analysis are specified.

For the developers, the use of the R package unity containing all functionalities and part of the workflow can facilitate code exchange and lower the possible errors during code writing. During package use and development you might encounter possible extensions - depending on specific requests of the users - which can be implemented and easily incorporated in previous as well as future reports. Although the development of such an R 'template' package might seem time-consuming at first, a lot of time is gained when using this package afterwards and making this effort worthwhile. For the users, reports are consistent across different analyses and appropriate package versioning - to keep track of changes, extensions and bug fixes - ensures the reproducibility of an entire analysis.

The **knitr** package can be used for the creation and successive integration of child templates.

A **shiny** app can be created to allow for an user-friendly and easy way of creating reports without need to be familiar to *R*.

Example of the implementation and the use of such workflow in **rmarkdown** format will be presented.

Simulate phenotype(s) with epistatic interactions

Beibei Jiang[1], Benno Pütz[1] and Bertram Müller-Myhsok[1]

1. Max Planck Institute for Psychiatry, Munich, Germany

Keywords: simulation, multiple phenotyes, epistatic interactions

Webpages: https://CRAN.R-project.org/package=SimPhe

For complex traits, genome-wide association studies (GWAS) are the standard tool to detect variants contributing to the variance of the phenotype of interest. However, limited to single-locus effects they can only explain a small fraction of the heritability of complex traits. Epistasis, generally defined as the interaction between different genes, has been hypothesized as one of the factors contributing to missing heritability. This has been a hot topic in quantitative genetics for a long time and there is a controversy about the role of epistasis because the majority of researchers only concentrate on additive effects as most genetic variation is (approximately) additive. Even for epistasis analysis, many tools cannot take the dominance effects into consideration properly. Recently, the detection of dominance or the interactions it is involved in have been reported. Meanwhile, simulation tools have been developed for evaluating type I error rates for new statistical association tests or power comparisons between the new tests and other existing tests. However, few of them focus on the dominance effect and its interactions with other genetic items. Here, we present an *R* package, **SimPhe**, to simulate single or multiple quantitative phenotypes based on genotypes with additive, dominance and epistatic effects using the Cockerham epistasis model. With optional parameters in different functions, users can easily specify the number of quantitative trait loci (QTLs), genetic effect size, the number of quantitative traits, and proportions of variance explained by the QTLs.

References

Cockerham, C. Clark, and Bruce Spencer Weir. 1977. "Quadratic Analyses of Reciprocal Crosses." *Biometrics* 33 (1). JSTOR: 187–203. doi:10.2307/2529312.

Gibran, Hemani, Shakhbazov Konstantin, Harm-Jan Westra, Tonu Esko, Anjali K. Henders, Allan F. McRae, Jian Yang, et al. 2014. "Detection and Replication of Epistasis Influencing Transcription in Humans." *Nature* 508 (April). Nature Publishing Group: 249–53. doi:10.1038/nature13005.

Kao, Chen-Hung, and Zhao-Bang Zeng. 2002. "Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model." *Genetics* 160 (3). Genetics Society of America: 1243–61. doi:10.1534/genetics.104.035857.

graphiT: an interactive, user-friendly tool to produce graphics based on the grammar of graphics' principles

VAUDOR Lise[1]

1. Université de Lyon, UMR 5600 EVS, ISIG

Keywords: ggplot2, shiny, graphics

Webpages: https://analytics.huma-num.fr/Lise.Vaudor/graphiT/

The package **graphiT** is based on both packages **shiny** and **ggplot2** and provides a user-friendly interface that helps users produce statistical graphics.

It is also a pedagogical tool, that helps with the teaching of the **ggplot2** package principles and use. Indeed, besides the graphic itself, it also provides the *R* command lines that would generate it, based on user input. Hence, **graphiT** is not only intended for users who are *R* and/or **ggplot2** newbies, but also for users who need a quick tool or reminder of the **ggplot2** commands.

graphiT is useable online here and also available as a gitHub repository here.

rOpenGov: community project for open government data

Lahti L[1], Lehtomäki J[2], Parkkinen J, and Kainu Mˆ3

1. Department of Mathematics and Statistics, University of Turku, Finland
2. Department of Earth Sciences, VU Amsterdam
3. The Social Insurance Institution of Finland

Keywords: Community Project, Infrastructure, Open Data, Open Science, R packages

Webpage: https://ropengov.github.io

Dedicated developer communities provide essential resources for the *R* ecosystem in the form of software packages, documentation, case studies, web applications, and other material. In 2010, we initiated the rOpenGov project (Lahti et al. 2013) to develop open source tools for open government data, computational social sciences, and digital humanities. We are a community of independent *R* package developers working on these topics in both public and private sector. Whereas the overall rOpenGov infrastructure is maintained by a core team, a number of independent authors have contributed projects and blog posts, supporting the overall objectives of this community-driven project. The main focus of our community is on knowledge sharing and peer support, and this has led to the release of several CRAN packages, including for instance the mature **eurostat** (Lahti et al. 2017), **pxweb**, and **gisfin** packages, and altogether over 20 projects at various stages of development and thousands of downloads per month. We welcome new contributions to the rOpenGov blog and the package collection, and participation on our online communication channels. In exchange, we aim to support the online community, provide feedback and support for package developers and initiate collaborations. Further details, full list of contributors, and up-to-date contact information are provided at the project website.

References

Lahti, Leo, Janne Huovari, Markus Kainu, and Przemyslaw Biecek. 2017. "Eurostat R Package." *R Journal. Accepted for Publication.* http://ropengov.github.io/eurostat.

Lahti, Leo, Juuso Parkkinen, Joona Lehtomäki, and Markus Kainu. 2013. "rOpenGov: open source ecosystem for computational social sciences and digital humanities." Presentation at Int'l Conf. on Machine Learning - Open Source Software workshop ICML/MLOSS). http://ropengov.github.io.

useR! 2017 Lightning Talk

Advanced R Solutions – A Bookdown Project

Malte Grosser [1,2], *Henning Bumann* [3], *Robert Krzyzanowski* [4], *Peter Hurford* [5]

1. *Department of Diagnostic and Interventional Neuroradiology, University Medical Center Hamburg-Eppendorf*
2. *implexis analytics*
3. *iqast.de*
4. *Avant Chicago*
5. *Charity Science Chicago*

Keywords: Advanced R, bookdown

Webpages: https://bookdown.org/Tazinho/Advanced-R-Solutions/

Working through a complex textbook can be cumbersome and frustrating. Despite a strong motivation to understand the content, one needs a good memory to bear all the details in mind, discipline to stay focused on the content, as well as patience to finish.

Solving exercises throughout the textbook can help to practice the learned, get more involved and gain deeper insights. Furthermore it can help to validate your newly gotten skills and understandings.

In this lightning talk, we will briefly discuss our experiences, while working through Hadley Wickham's Advanced R book (Wickham 2014), which provides exercises after most of its chapters. In particular we will describe the approach to document and monitor our progress via Yihui Xie's **bookdown** package (Xie 2017) to address the issues mentioned above.

References

Wickham, Hadley. 2014. *Advanced R*. Boca Raton, Florida: Chapman; Hall/CRC. http://adv-r.had.co.nz/.

Xie, Yihui. 2017. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.

Shiny Apps for Maths and Stats Exercises

Marlene Müller[1]

1. Beuth University of Applied Sciences Berlin, Germany

Keywords: shiny apps, student exercises

Webpages: http://prof.beuth-hochschule.de/mmueller/shiny-apps/

The talk presents some ideas to generate exercises for Maths and Stats courses using shiny apps. The *R* software environment in the background is quite useful here: It allows to randomly choose parameters and data in such exercises and provides calculation results and graphical illustrations. In addition, the MathJax capability of shiny allows to use formulas as in classical textbooks.

Since shiny apps generate web documents, they can be easily linked into websites or online platforms (e.g. into Moodle courses). Self-written shiny apps allow for an implementation that precisely fits the needs of a specific mathematics or statistics course.

The Maths and Stats apps presented here are intended as a complementary offer, i.e. in addition to usual classroom exercises. The students should use them independently to train and self-test their mathematics or statistics skills.

R in research on microbial mutation rates

Qi Zheng

Texas A&M School of Public Health, College Station
qzheng@sph.tamhsc.edu

The determination of microbial mutation rates in the laboratory is a routine yet computationally challenging task in biological research. The experimentalist conducts experiments in accord with the classic Luria-Delbrück protocol [1] (aka the fluctuation assay protocol). But the resulting fluctuation assay data can pose a formidable challenge, not only to bench biologists, but also to bioinformaticians unfamiliar with the biological and statistical subtleties inherent in the fluctuation assay protocol. Due to the increasing popularity of the fluctuation experiment in recent biological research, more and more bench scientists are eager to analyze their fluctuation assay data by themselves. Some understandably expect the analyses to be as simple as calculating the sample mean using a pocket calculator. The popular web tool FALCOR [2] almost fulfilled this dream. Unknown to most practitioners, this web tool has important limitations, which can lead an unwary user to faulty conclusions [3]. For example, the comparison of microbial mutation rates is beyond the capabilities of this web tool. The R package **rSalvador** (available at http://eeeeeric.github.io/rSalvador) makes accessible to bench scientists a myriad of latest computational methods that are necessary for proper analysis of fluctuation assay data. **rSalvador** allows the user to properly account for relative fitness and plating efficiency. In particular, **rSalvador** is the only tool that affords strictly likelihood-based methods for the comparison of microbial mutation rates. This presentation will discuss the role of **rSalvador** in biological education, in mutation research, and in the development of new algorithms for fluctuation assay data.

References

[1] SE Luria, M Delbrück (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491-511.

[2] BM Hall, C-X Ma, P Liang, KK Singh (2009) Fluctuation AnaLysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbrück fluctuation analysis. Bioinformatics 25:1564- 1565.

[3] Q Zheng (2015) A new practical guide to the Luria-Delbrück protocol. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 781:7-13.

DNA methylation-based classification of human central nervous system tumors

Martin Sill [1,2], Volker Hovestadt [3], Daniel Schrimpf [4], David Jones [1], David Capper [4,5], Stefan Pfister [1] and Andreas von Deimling [4,5]

1. Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
2. Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany
3. Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany
4. Department of Neuropathology, University Hospital Heidelberg, Heidelberg, Germany
5. Clinical Cooperation Unit Neuropathology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Keywords: machine learning, bioinformatics, methylation data, brain tumor diagnosis

Webpages: https://www.molecularneuropathology.org/mnp

More than 100 brain tumor entities are listed in the World Health Organization (WHO) classification. Most of these are defined by morphological and histochemical criteria that may be ambiguous for some tumor entities and if the tissue material is of poor quality. This can make a histological diagnosis challenging, even for skilled neuropathologists. Molecular high-throughput technologies that can complement standard histological diagnostics have the potential to greatly enhance diagnostic accuracy. Profiling of genome-wide DNA methylation patterns, likely representing a 'fingerprint' of the cellular origin, is one such promising technology for tumor classification.

We have collected brain tumor DNA methylation profiles of almost 3,000 cases using the Illumina HumanMethylation450 (450k) array, covering over 90 brain tumor entities. Using this extensive dataset, we trained a Random Forest classifier which predicts brain tumor entities of diagnostic cases with high accuracy (Capper et al. 2017). 450k methylation data can also be used to generate genome-wide copy-number profiles and predict target gene methylation. We have developed a *R* package that includes a data analysis pipeline which takes data of the Illumina 450k array and the new EPIC array as input and automatically generates diagnostic reports containing quality control metrics, brain tumor class predictions with tumor class probability estimates, copy number profiles and target gene methylation status.

Besides sharing this *R* package with cooperating institutes worldwide, we also offer a web interface that allows researchers from other institutes to apply the pipeline to their own data. Practical experience from different cooperating institutes show that application of our pipeline to Illumina methylation array data represents a cost efficient method to greatly improve diagnostic accuracy and clinical decision making.

References

Capper, David, David Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, . . . , Andreas von Deimling, and Stefan Pfister. 2017. "DNA Methylation-Based Classification of Human Central Nervous System Tumors." *Submitted to Nature.*

useR! 2017 Lightning Talk

redmineR and the story of automating *useR!2017* abstract review process

Maxim Nazarov[*]

Keywords: Redmine, project management, API, automation, useR!

Redmine is a popular open-source project management platform with rich capabilities. Luckily it has a RESTful API with bindings for many common programming languages, alas not including *R* until now. We introduce **redmineR**, an *R* interface to Redmine API to close this gap.

As a use case we give a glimpse into useR! conference organizer's life. The *useR!2017* abstract revision process was automated using *R* and leveraging the capabilities of Redmine with **redmineR**. Abstracts submitted were extracted from a MySQL database used for the web site back-end and converted into Redmine "issues" with **redmineR**, automatically split into sub-projects by topics. Reviewers for each topic having access to the corresponding sub-project were automatically notified by Redmine. All communication for the review process also happened inside Redmine with reviewers communicating through comments and assignments. After the review process **redmineR** was used to poll all abstracts with "Accepted" status to send automatic email notifications to the lucky authors.

[*] Open Analytics, Belgium

What is missing from the meta-analysis landscape?

Michael Dewey[1]

1. King's College London UK

Keywords: meta-analysis, CRAN Task View

At the time of writing the CRAN Task View MetaAnalysis contains 70+ packages. Despite this large resource there are many overlaps and some important gaps. After a very brief overview of meta-analysis we shall discuss the relationships between the existing packages and the coverage they offer. The main techniques for preparing summary statistics and analysing them in univariate models are well covered with considerable overlaps. Graphical displays are another area of strength but diagnostics have received less coverage. Other topics with adequate coverage include meta-analysis of diagnostic tests, multivariate meta-analysis and network meta-analysis. We shall then outline the current gaps: meta-analysis where some studies have individual participant data available but others only have summary statistics (discussed in the literature but not coded), the method of Hunter and Schmidt (primarily covered in books with closed source software), and trial sequential analysis (discussed in the literature and available in closed source software).

Introducing the DynNom package for the generation of dynamic nomograms

John Newell[1], Amirhossein Jalali[1], Davood Roshan[1] and Alberto Alvarez-Iglesias[2]

1. School of Mathematics, Statistics and Applied Mathematics, NUI Galway
2. HRB Clinical Research Facility, NUI Galway

Keywords: Nomograms, Shiny, Dynamic Nomograms, Translational Statistics

Webpages: https://cran.r-project.org/web/packages/DynNom/index.html

Nomograms are useful computational tools for model visualisation as they allow the calculation of a point estimate of a response variable for a set of values of the corresponding explanatory variables. The `nomogram` function in Frank Harrell's **rms** package is a popular way of creating static nomograms for regression models. Our **DynNom** package, built using **shiny**, allows the creation of dynamic nomograms, using a simple wrapper function from a variety of model objects including `lm`, `glm` and `coxph` and models generated using the `Ols`, `Glm`, `lrm` and `cph` functions in the **rms** package. In this presentation examples will be given where dynamic nomograms will be generated for a variety of models and the potential of this approach to be a useful translational tool explored.

useR! 2017 Lightning Talk

Better Confidence Intervals for Quantiles

Michael Höhle[1]

1. Department of Mathematics, University of Stockholm, Sweden

Keywords: Coverage, Comparing R packages, Uncertainty, Bootstrap

Webpages: http://www.math.su.se/~hoehle

Inspired by the work of Höhle and Höhle (2009) concerned with the assessment of accuracy for digital elevation models in photogrammetry, we discuss the computation of confidence intervals for the median or any other quantile in R. In particular we are interested in the interpolated order statistic approach suggested by Hettmansperger and Sheather (1986) and generalized in Nyblom (1992). In order to make the methods available to a greater audience we provide an implementation of these methods in the R package `quantileCI` and conduct a small simulation study to show that these intervals indeed have a very good coverage. The study also shows that these intervals perform better than the currently available approaches in R. We therefore propose that these intervals should be used more in the future!

Details on the work can be found in the presenter's blog entitled *Theory meets practice* available at http://www.math.su.se/~hoehle/blog.

References

Hettmansperger, T. P., and S. J Sheather. 1986. "Confidence Intervals Based on Interpolated Order Statistics." *Statistics and Probability Letters* 4: 75–79. doi:10.1016/0167-7152(86)90021-0.

Höhle, J., and M. Höhle. 2009. "Accuracy Assessment of Digital Elevation Models by Means of Robust Statistical Methods." *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (4): 398–406. doi:10.1016/j.isprsjprs.2009.02.003.

Nyblom, J. 1992. "Note on Interpolated Order Statistics." *Statistics and Probability Letters* 14: 129–31. doi:10.1016/0167-7152(92)90076-H.

An example of Shiny tool at Nestlé R&D, an enabler to guide product developers in designing gluten free biscuits

M.Lepage[1], A.Moroni[1], R-L.King[1], C.Gancel[1], S.Faneco[1] and S.Boudalier[2]

1. Nestlé Research Center, Lausanne, Switzerland
2. Nestlé Development Center, Santiago, Chile

Keywords: Shiny, linear modeling, gluten-free, formulation guidance tool

What are the expected differences in texture while comparing two recipe of gluten free (GF) biscuits that have different ingredients (type and/or quantity)? A shiny tool has been developed to address this question.

Knowing that the wheat flour components: gluten proteins, but also starch and fiber are responsible for the behavior of the formed dough as well as of the baked product, the main challenges of developing GF recipes deal therefore with controlling the final texture of the product. To compensate the absence of wheat flour, combinations of GF flours, starches, proteins and fibers having different functional properties have been tested in order to create a diversity of different texture of the final product, while ensuring the right nutritional intake. 32 biscuits have been produced following a design of experiments that combine the four groups of ingredients in a balanced way (both in terms of quality and quantity). A sensory trained panel have assessed the organoleptic properties of these biscuits (texture, appearance, flavor, taste)

This tool displays an interactive visualization of statistical linear models that relate the key functional properties of the ingredients (X) and the texture characteristics of biscuits (Y).

Quantifying the impact of the recipe on product properties and thereby helping to reduce the number of trials needed to optimize the final product is a key element for innovation and renovation within Nestl?. Complementing the knowledge of product developers by providing them Shiny guidance tools is definitely the direction to go.

Using R to Analyze Healthcare Cost of Older Patients Using Personal Emergency Response Service

M. Simons[1], S. Golas[2], J op den Buijs[1], N. Fischer[2], J. Felsted[2], L. Schertzer[3], S Agboola[2]
1. Philips Research
2. Partners Connected Health
3. Philips Lifeline

Keywords: Older patients, Personal Emergency Response Service (PERS), Healthcare cost analysis

Webpages: Partners Connected Health, Philips Lifeline

Introduction

In the US healthcare system, half of overall Medicare and Medicaid services reimbursement goes towards caring for the top 5% most expensive patients. However, little is known about patient and cost flow prior to reaching the top 5% and how dynamic these flows are from year to year. To address this gap we analyzed patient flow and associated healthcare cost trends over a five year period.

Methods

This is a retrospective, longitudinal, multicenter study to evaluate healthcare cost of 2,642 older patients over the period 2011-2015. The study population was segmented by their healthcare expenditure into Top- (5%), Middle (6-50%) and Bottom-(51-100%) segments to build cost acuity pyramids for each fiscal year. The longitudinal healthcare expenditure trends of the complete study population as well as each segment were assessed by linear regression models. Patient flows throughout the segments of the cost acuity pyramids from year to year were modeled by Markov chains. The associate costs flows were quantified over a 2-year period. All statistical analysis was performed in *R* using packages such as **data.table**, **dplyr**, **stats**, **ggplot2**.

Results

Total healthcare cost of our study population nearly doubled from $17.7 M in 2011 to $33.0 M in 2015, although the number of patients per year having any costs was steady. The increasing trend in total healthcare cost is statistically significant ($p = 0.003$) and the expected cost increase every next year is $3.6 M. The majority of this increasing trend was contributed to the M-segments with $2.3 M expected cost increase ($p = 0.002$), followed by the T- and B-segments with $1.2 M and $0.1 M expected cost increase, respectively ($p = 0.008$ and $p = 0.003$). The patients and cost flow analyses showed that 18% of patients moved up the cost acuity pyramid yearly and their cost increased by 672% in contrast to 22% of patients that moved down with a cost decreased by 86%. The remaining 60% of patients stayed at the same segment next year, however their cost increased by 18%.

Conclusions

While currently healthcare organizations target the most expensive patients (top 5%) by high-cost intensive care programs, our patient and cost flow analysis unveils cost savings opportunities by providing interventions to the patients in the lower segments that are moving up the cost acuity pyramid. Inexpensive, senior accepted technology such as Personal Emergency Response Service that continuously monitors patients at home in real-time could help identify seniors who need targeted care, thereby providing opportunities for cost savings, high quality care, and improved patient satisfaction.

Digital Signal Processing with R

Munshi Imran Hossain[1], Andrea Hita[2] and Rajat Mukherjee[3]

1. Software Affiliate, Cytel Statistical Software and Services Pvt. Ltd., Pune, India
2. Data Scientist, Cytel, Spain
3. Biostatistician, Cytel, Spain

Keywords: Signal processing, signal segmentation, predictive modeling

Digital signal processing (DSP) is widely done using languages like *MATLAB* and *C++*. *R* is relatively less known for its support in this domain. However *R* provides strong support for DSP work.

In this talk, we will demonstrate the solution to a standard signal processing and segmentation problem using R. In this problem we will use an input stream of data such as a time series, that will be analysed for power content using short-term Fourier transform methods. The power content in the frequency range of interest, is then smoothed to attenuate short-term variations. This can then be suitably thresholded to find the segment of the signal that corresponds to the region of interest.

The implementation will be done purely in *R* using packages such as, **signal**, **zoo**, **e1071**, etc. Visualisation of the processed data at intermediate stages is also useful and this can be done very efficiently using packages such as **ggplot2**.

Ultra-Fast Data Mining With The R-KDB+ Interface

Nataraj Dasgupta

RxDataScience Inc.
nd@rxdatascience.com

Keywords: kdb+, big data mining, R-KDB+ Interface, business/industry, high-performance computing

Webpages: http://code.kx.com/wiki/Cookbook/IntegratingWithR

Commercial application of ultra-low latency techniques for data mining and machine learning have been ubiquitous in financial trading and related disciplines for many years. As early as 2005, algorithmic trading desks at hedge funds and large investment banks have relied on in-memory, columnar databases and map-reduce techniques for analysing millions of data points in milliseconds long before such tools were used in other verticals. In particular, technologies such as *kdb+* and *Q* - a vector-based programming platform developed as a successor to *APL* (*A Programming Language* developed in the 1950s/60s in Harvard by Ken Iverson), provide an unchallenged ability to perform both simple and complex data manipulations at scale with speeds that are orders of magnitude faster than contemporary platforms used for Big Data. A lesser-used, but formidable capability that has been used by *R*-enthusiasts who were also *kdb+* experts has been the *R-KDB+* Interface used for interprocess-communication to share data between *R* and *KDB+* processes all from within the user's *R*-console or *Q*-console. In my nearly, 12 years of using *R*, I, like many of my colleagues who have worked in financial trading environments have found such capabilities indispensable especially when working with large, oftentimes, TeraByte-scale datasets. The proposed talk features the basics of using the *R-KDB+* interface as a faster, superior and more optimal method to extract aggregated data from TB-scale data warehouses prior to statistical analysis in R.

Strategies for Reproducible Research with Packrat

Sean Lopp[1]

1. RStudio

Keywords: Reproducible Research, Packrat, Packages, Version Control

Webpages: https://CRAN.R-project.org/package=packrat, https://rstudio.github.io/packrat/

packrat is a dependency management tool for R. Though **packrat** was released three years ago, many R users have struggled to approach and adopt the tool. This talk will highlight different use cases for **packrat** and corresponding strategies for success. Demos will include sharing code with others, moving code between environments, and tracking project dependencies alongside of code in version control. We'll demonstrate how **packrat** works traditionally, but also explain an alternative workflow used in production everyday at RStudio.

TAGS - Table Assorting Guided System: an HTML widget to create multiple tables from Excel spreadsheets

Paulo R. Bargo and Davit Sargsyan

Statistical and Decision Sciences, Janssen Research and Development LLC

Keywords: Excel spreadsheet, htmlwidget, data import

Despite advancements in data storing and sharing which allows for direct access to database content though API calls (i.e., SQL, NoSQL, etc.), Microsoft Excel™ and other spreadsheet applications remain widespread as data entry, distribution and presentation platforms for scientists across all fields of research and development. Spreadsheet applications have obvious advantages in terms of intuitive representation of data as tables, ease of use, and minimal to no programming effort. However, they generally have limited graphing and analytical capabilities. Adding large number of formulas and interactive plots can significantly slow down the application. Up until now, most attempts to extend Excel's capabilities have consisted of connecting an Excel interface to an analytical engine, usually R, using Visual Basics for Excel (VBA) programming language. While it allows for relatively seamless integration of the two, the downside of this solution is that it requires proficiency in both VBA and R programming to develop new procedures. We have decided to take an alternative approach and bring Excel into R. Using Shiny® technology, we are able to import, display and interact with Excel workbooks inside web applications. An immediate benefit of this approach is that all programming can now be done in R without ever resorting to VBA. This opens doors to large number of possibilities, from flexible selection of data ranges to new methods of storing and retrieving data. Spreadsheets typically are constructed with additional metadata, making them very convoluted. It is also common to have multiple tables included in a single spreadsheet separated by empty columns or rows, which complicates the import of data to an R session. The arduous task of "cleaning" the files normally falls to the data scientist, who must either hard code the exact location of the tables in the spreadsheets using customized commands to retrieve the data, or allocate it to separate CSV files for posterior use. This last option should always be avoided, due to the possibility of data duplication. TAGS - Table Assorting Guided System, is an (JS-based) htmlwidgets package created to simplify the work of the data scientist while retrieving data from complex Excel spreadsheets. The package loads the data into a webpage display configuration similar to that of Excel. The user can then click and drag the mouse to highlight the location of the data in the spreadsheet, and tag it. Basic information about the table, such as the file location, name and sheet, is automatically added to the tag metadata. The user can than add any number of key:value pairs by typing them into a dialog box therefore expanding the number of table descriptors. The metadata is saved in a JSON file that can be used later to retrieve the data. Simultaneously, an S4 object containing both the metadata and values from the selected spreadsheet range are available to current R session. We will discuss the creation of the package as well as demonstrate its use. TAGS will reduce time lost and frustration when data are imported from spreadsheets, eliminate copying/pasting, and improve reproducibility.

useR! 2017 Lightning Talk

The current state of naming conventions in R.

Rasmus Bååth[1] and Xavier Guardiola[1]

1. King.com Ltd. (Activision Blizzard)

Keywords: Naming Conventions, R

Coming from another programming language one quickly notes that there are many different naming conventions in use in the *R* community. Looking through packages published on CRAN one will find that functions and variables most often are either `period.separated` or `underscore_separated`, or written in `lowerCamelCase` or `UpperCamelCase`. In 2012 we did a survey of the popularity of different naming conventions used in all the packages on CRAN (Bååth, 2012), but a lot has happened since then! Since 2012 CRAN has more than doubled from 4000 packages to now over 10,000 packages, and we have also seen the rising popularity of the **tidyverse** packages that often follow the `underscore_separated` naming convention.

In this presentation we will show you the current state of naming conventions used in the R community, we will look at what has happened since 2012 and what the current trend is.

References

Bååth, R. (2012). The state of naming conventions in R. *The R Journal*, 4(2), 74-75. https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Baaaath.pdf

R in a Pharmaceutical Company

Reinhold Koch

Motivation

Initiated by recently graduated statisticians a push for R is noticeable. Nice development tools and more so shiny web-based reports attract interest from management.

Implementation

From a company perspective some technical foundation for R has to be provided beyond user's desktop systems, for instance dedicated server(s) and maybe some High Performance Cluster - and all to interact nicely with each other.

Maintenance

The R ecosystem evolves at a good pace. When to centrally install which R version, which packages can raise a number of issues.

Infrastructure

Besides R and shiny servers we found generally web portals were appreciated by users. For R code development a central git repository server seems indispensable nowadays, best with facilities for continuous integration and a server for locally developed packages.

Community

In a company spread out across many time zones electronic support for discussions about R helps a lot to keep the ball rolling and advance usage. This is no substitute for local meetings that easily can span across department boundaries.

Similarly training in R benefits from a two-pronged strategy:

- electronic
  via self-paced online courses
- local classroom
  for dedicated topics
- peer communication

Jurimetrics: quantitative analysis of judicial decisions using *R*

José de Jesus Filho[1] *and Julio Adolfo Zucon Trecenti*[2]

1. Fundação Getulio Vargas
2. Universidade de São Paulo

Keywords: Jurimetrics, Judicial decisions,Webscraping,Text mining, Predictive modeling

Webpages: https://josejesus.info

Abstract: The increasing availability of online access to judicial decisions coupled with modern *R* packages that perform webscraping, text mining, topic modeling and predictive modeling, allow for the application of quantitative methods in the simultaneous analysis of thousands of court judgements. Extraction, manipulation, and analysis of judicial decisions require a variety of tecniques and the use of multiple *R* packages, as well as bulding new functions to attain and analyze relevant content. As an example, unsupervised learning, such as topic modeling, has revealed here-to unknown aspects of how courts handle traffic accident cases. Supervised learning, such as classification, is a very important tool to identify determinants of judicial decisions, which are influenced by the interpretation of facts and, suprisingly, by judges' ideology. It is now possible to predict with a high level of accuracy how courts will decide in criminal cases. The talk will address a set of tecniques that have been developed to analyze court rulings.

useR! 2017 Lightning Talk

Object-oriented markdown in R to facilitate collaboration

Rickey E. Carter and Tina M. Gunderson

Department of Health Sciences Research, Mayo Clinic, Rochester, MN USA

Abstract: A frequently expressed barrier to the transition from SAS to R at our institution is the challenge in generating "quick and dirty" output that combines text and graphical summaries of data for offline viewing or sharing with investigators. Depending on a person's prior training and programming style, a full markdown approach to produce this integrated summary often requires significant reprogramming, particularly when the project involves multiple programmers or complex data manipulation. The philosophy behind an approach entitled "object-oriented markdown" will be presented and illustrated using a series of research projects utilizing the **RJafroc** package. The presentation will illustrate how data management and analysis standards can provide a framework that enables collaboration amongst statisticians on the project and ease of integration of final statistical results into a markdown document. By utilizing a markdown file only as a means to print stored R objects, one is able to rapidly summarize and interpret statistical output while maintaining efficient programming styles.

Keywords: Reproducible research, statistical output, data analysis pipeline

useR! 2017 Lightning Talk

The **dataCompareR** package

Rob Noble-Eddy[1]

1. Capital One UK

Keywords: packages, finance, data

Webpages: http://www.capitalone.co.uk, https://github.com/capitalone

At Capital One (UK) the Data Science team use R for almost all analytics, making use of a wide range of existing packages from the CRAN. When we identified some extra functionality that was needed, we were keen to build it as package to give back to the open source community. The team built **dataCompareR** towards the end of 2016, and after a few months of internal testing, will make the package available on the CRAN, becoming our first package contributed to the R community.

dataCompareR aims to make it easy to compare two tabular data objects in R. It's specifically designed to show differences between two sets of data in a useful way that should make it easier to understand the differences, and if necessary, help you work out how to make them match. It aims to offer a more useful output than `all.equal` when your two datasets do not match, but isn't intended to replace `all.equal` as a way to test for equality.

In this talk I will briefly cover our experience of transitioning to R and the way we designed and built **dataCompareR**, before focusing on the functionality of the package.

minimalRSD and FMC: R packages to construct cost efficient experimental designs

Shwetanki Lall, Seema Jaggi, Eldho Varghese, Arpan Bhowmik and Cini Varghese

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Keywords: experimental designs, run sequences, cost effective

In this talk, we explore and discuss the possibility of reducing overall cost and effort in a scientific experiment. According to the objectives and assumptions of the study, an experimenter can adopt a suitable experimental design using available numerous tools, software and R packages. But such designs do not consider the sequence of experimental runs to be applied on experimental units. This might result in increase in cost and effort, e.g., if a factor in the experiment is temperature, then the experimenter might have to change the temperature levels from high to low many times in successive runs and in doing so he/she has to wait and adjust the instrument many times. We have addressed this issue and proposed theoretical framework for minimizing the changes in factor levels in an experimental design. To apply our findings we developed two R packages: minimalRSD and FMC. Package minimalRSD can be used to generate response surface designs namely, central composite designs (CCD) with full factorial or fractional factorial points and Box Behnken designs (BBD) and the factorial designs with symmetrical as well as asymmetrical factor level combinations can be constructed using the package FMC. The output gives the respective design, the number of changes in each factor and the overall number of level changes. We intend to extend our theoretical findings to the scientific community using the power of R.

References

Eldho Varghese, Arpan Bhowmik, Seema Jaggi, Cini Varghese and Shwetank Lall (2017).

On the construction of response surface designs with minimum level changes Utilitas Mathematica (Under print).

Arpan Bhowmik, Eldho Varghese, Seema Jaggi and Cini Varghese (2016). Minimally changed run sequences in factorial experiments Communications in Statistics—Theory and Methods. doi:10.1080/03610926.2016.1152490.

ICtest: Estimating the Number of Interesting Components

Klaus Nordhausen[1], Hannu Oja[1] David E. Tyler[2] and Joni Virta[1]

1. *Department of Mathematics and Statistics, University of Turku*
2. *Department of Statistics, The State University of New Jersey*

Keywords: Principal component analysis, Independent component analysis, Non-Gaussian component analysis, Sliced inverse regression

Webpages: https://CRAN.R-project.org/package=ICtest

Choosing the number of components to retain is a crucial step in every dimension reduction method. The package **ICtest** introduces various tools for estimating the number of interesting components, or the true reduced dimension, in three classical situations: principal component analysis, independent component analysis and reducing the number of covariates in prediction. The estimation methods are provided in the form of hypothesis tests and in each of the three cases tests based both on asymptotic distributions and on bootstrapping are provided. The talk goes to shortly introduce the used methodology and showcase the package in action.

References

Nordhausen, Klaus, Hannu Oja, and David E. Tyler. 2016. "Asymptotic and Bootstrap Tests for Subspace Dimension." *arXiv:1611.04908*.

Nordhausen, Klaus, Hannu Oja, David E. Tyler, and Joni Virta. 2017. "Asymptotic and Bootstrap Tests for the Dimension of the Non-Gaussian Subspace." *arXiv:1701.06836*.

Virta, Joni, Klaus Nordhausen, and Hannu Oja. 2016. "Projection Pursuit for Non-Gaussian Independent Components." *arXiv:1612.05445*.

useR! 2017 Lightning Talk

heatmaply: an *R* package for creating interactive cluster heatmaps

Tal Galili[1]

1. Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences. Tel Aviv University.

Keywords: cluster heatmap, interactive visualization, ggplot2, plotly, shiny

Webpages: heatmaply, shinyHeatmaply

A cluster heatmap is a popular graphical method for visualizing high dimensional data, in which a table of numbers are encoded as a grid of colored cells (Wilkinson and Friendly 2009, Weinstein (2008)). The rows and columns of the matrix are ordered to highlight patterns and are often accompanied by dendrograms and extra columns of categorical annotation. Heatmaps are used in many fields for visualizing observations, correlations, and missing values patterns. There are many *R* packages and functions for creating static heatmap figures (the most famous one is probably `gplots::heatmap.2`).

The **heatmaply** R package allows the creation of interactive cluster heatmaps, enabling tooltip hover text and zoom-in capabilities (from either the grid or the dendrograms), while supporting sidebar annotation. The package brings together many well known packages such as **ggplot2** (Wickham 2016), **plotly**, **viridis**, **seriation** (Hahsler, Hornik, and Buchta 2008), **dendextend** (Galili 2015), and others. Also, it is now supported by the **shinyHeatmaply shiny** app.

You can play with a simple interactive example by running:

```
install.packages('heatmaply'); library('heatmaply')
heatmaply(percentize(mtcars), k_row = 4, k_col = 2, margins = c(40,120,40,20))
```

This talk will provide an overview of design principles for creating a useful, and beautiful, cluster heatmap. Attention will be given to data preprocessing, choosing a color palette, and careful dendrograms creation.

This work was made possible thanks to the essential contribution of Jonathan Sidi, Alan O'Callaghan, Carson Sievert, and Yoav Benjamini. As well as the joint work of Joe Cheng and myself on the **d3heatmap** package (which laid the foundation for **heatmaply**). The speaker is the creator of the R packages **installr**, **dendextend**, and **heatmaply**, and blogs at: www.r-statistics.com.

References

Galili, Tal. 2015. "Dendextend: An R Package for Visualizing, Adjusting and Comparing Trees of Hierarchical Clustering." *Bioinformatics*. Oxford Univ Press, btv428.

Hahsler, Michael, Kurt Hornik, and Christian Buchta. 2008. "Getting Things in Order: An Introduction to the R Package Seriation." *Journal of Statistical Software* 25 (3). American Statistical Association: 1–34.

Weinstein, John N. 2008. "A Postgenomic Visual Icon." *Science* 319 (5871). American Association for the Advancement of Science: 1772–3.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wilkinson, Leland, and Michael Friendly. 2009. "The History of the Cluster Heat Map." *The American Statistician* 63 (2). Taylor & Francis: 179–84.

Multivariate statistics for PAT data analysis: short overview of existing R packages and methods

Tatsiana Khamiakova[1], Nicolas Sauwen[2], Michel Thiel[1], Tor Maes[1], Helena Geys[1]
1 Janssen Pharmaceutica
2 Open Analytics

Keywords: chemoinformatics, multivariate data analysis, time series data

Process analytical technology (PAT) is defined as a system for designing, analyzing, and controlling pharmaceutical manufacturing processes through timely measurements (i.e., during processing) of critical quality and performance attributes. PAT poses many data analysis challenges, such as appropriate techniques for data preprocessing, quantification and integration with the external information (e.g. DoE factors). Currently available *R* packages in the public repositories allow for integrated analysis and implementation of analytic pipelines in the industrial setting. In this presentation we will focus on a specific application of using infrared (IR) spectroscopy technology for synthesis reaction monitoring and multivariate analysis of IR spectra by using matrix factorization techniques (e.g. principal component analysis, factor analysis for bicluster acquisition, non-negative matrix factorization, time series factor analysis and curve resolution). Unlike a supervised partial least squares technique - which is commonly used in chemometrics - this is a set of unsupervised techniques implemented in *R* which allow to extract and explore the most essential information in the IR spectra. The ability to extract this type of information reduces the task of monitoring about 600 highly correlated points per spectrum to monitoring a few independent factor scores only. For the proof of concept, the scores from several matrix factorization methods are compared to the known compound concentrations and the differences and commonalities of the different approaches are discussed.

useR! 2017 Lightning Talk

eseis – A toolbox to weld seismologic and geomorphic data analysis

Michael Dietze[1]

1. German Research Centre for Geosciences GFZ Potsdam

Keywords: Seismology, Spatial, Time series, Geomorphology

Webpages: http://micha-dietze.de/pages/eseis.html, https://github.com/coffeemuggler/eseis

Environmental seismology is the science of investigating the seismic signals that are emitted by Earth surface processes. This emerging field provides unique opportunities to identify, locate, track and inspect a wide range of the processes that shape our planet. Modern broadband seismometers are sensitive enough to detect signals from sources as weak as wind interacting with the ground and as powerful as collapsing mountains. This places the field of environmental seismology at the seams of many geoscientific disciplines and requires integration of a series of specialised analysis techniques.

R provides the perfect environment for this challenge. The package eseis uses the foundations laid by a series of existing packages and data types tailored to solve specialised problems (e.g., signal, sp, rgdal, Rcpp, matrixStats) and thus provides access to efficiently handling large streams of seismic data (> 300 million samples per station and day). It supports standard data formats (mseed, sac), preparation techniques (deconvolution, filtering, rotation), processing methods (spectra, spectrograms, event picking, migration for localisation) and data visualisation. Thus, eseis provides a seamless approach to the entire workflow of environmental seismology and passes the output to related analysis fields with temporal, spatial and modelling focus in R.

smires – Calculating Hydrological Metrics for Univariate Time Series

Tobias Gauster [1] *and Gregor Laaha* [1]

1. *Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, Vienna*

Keywords: Ecology, Hydrology, Framework, Time Series

Webpages: https://github.com/mundl/smires, http://www.smires.eu/

Many hydrological and ecological metrics are constructed in a similar way. A common family of metrics is calculated from a univariate time series (e.g. daily streamflow observations) aggregated for given periods of time. More complex ones involve the detection of events (e.g. no-flow periods or flood events) or several levels of aggregation (e.g. mean annual minimum flow).

Although some *R* packages (**hydrostats**, **IHA**, **hydroTSM**, . . .) providing hydrological metrics exist, they usually strictly require daily time series and do not allow for a free choice of the aggregation period. By contrast the package **smires** tries to generalize the calculation and visualization of hydrological metrics for univariate time series providing a generic framework which is developed around **dplyr**s (Wickham and Francois 2016) split-apply-combine strategy. It takes into account the peculiarities of hydrological data e.g., the strong seasonal component or the handling of missing data.

The general approach comprises four steps. (1) First the time series can be *preprocessed*, e.g. by interpolating missing values or by applying a moving average. If necessary, an optional step (2) involves the identification of *distinct events* such as low flow periods. For each event a set of new variables (e.g. event duration or event onset) is derived. In a third step (3) *summary statistics* are calculated for arbitrary periods (e.g. months, seasons, calendar years, hydrological years, decades). This step can be repeated until the original time series is aggregated to a single value.

The user keeps full control over the frequency of the time series (daily, weekly, monthly), the choice of preprocessing functions, the aggregation periods, the aggregation functions as well as the handling of events spanning multiple periods. Thus, **smires** enables the user to obtain a wide range of metrics whilst minimizing programming effort and error-proneness.

References

Bond, Nick. 2016. *Hydrostats: Hydrologic Indices for Daily Time Series Data*. https://CRAN.R-project.org/package=hydrostats.

The Nature Conservancy. 2009. *Indicators of Hydrologic Alteration*. https://www.conservationgateway.org/ConservationPractices/Freshwater/EnvironmentalFlows/MethodsandTools/IndicatorsofHydrologicAlteration/Pages/indicators-hydrologic-alt.aspx.

Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zambrano-Bigiarini, Mauricio. 2014. *HydroTSM: Time Series Management, Analysis and Interpolation for Hydrological Modelling*. https://CRAN.R-project.org/package=hydroTSM.

Candy Crush R Saga

Xavier Guardiola[1] and Rasmus Bååth[1]

1. King.com Ltd. (Activision Blizzard)

Keywords: big data, packages, shiny, reproducible research, business analytics

Abstract: Over the last 5 years mobile gaming industry has experienced a massive growth. Hundreds of millions of players play King games every month. We will briefly talk about how King data science teams coped with this amount of information and how **R** has allowed them to collaborate and grow a culture of reproducible research. Large data science teams are typically made of people coming from very different backgrounds: hard sciences, engineering, economics, computer science, business, psychology, etc. That poses the challenge to develop a common technology stack that could allow for a fluid and agile collaboration. **R** is the perfect language for that.

Namely, we will see how **R** has been used to:

- Build data access packages
- Quickly assemble dashboards and reporting tools by leverging the **shiny** package.
- Implement a reproducible research mindset with **github**, **Rmarkdown** and **notebooks**.

Using an *R* package as platform for harmonized cleaning of data from RTI MicroPEM air quality sensors

Maëlle Salmon[1], Sreekanth Vakacherla[2], Carles Milà[1], Julian D. Marshall[2], Cathryn Tonne [1]

1. ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Universitat Pompeu Fabra, CIBER Epidemiología y Salud Pública, Barcelona, Spain.
2. Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA

Keywords: package, reproducibility, science, quality control, personal monitoring

Webpages: http://www.masalmon.eu/rtimicropem/

RTI MicroPEM is a small particulate matter personal exposure monitor, increasingly used in developed and developing countries. Each measurement session produces a csv file which includes a header with information on instrument settings and a table of thousands of observations of time-varying variables such as particulate matter concentration, relative humidity. Files need to be processed for 1) generating a format suitable for further analysis and 2) cleaning the data to deal with the instruments shortcomings. Currently, this is not done in a harmonized and transparent way. Our package pre-processes the data and converts them into a format that allows the integration the rich set of data manipulation and visualization functionalities that the **tidyverse** provides.

We made our software open-source for better reproducibility, easier involvement of new contributors and free use, particularly in developing countries. We applied the package in a research project for a large number of measurements. The functionalities of our package are three-fold: allowing conversion of files, empowering easy data quality checks, and supporting reproducible data cleaning through documentation of current workflows.

For inspection of individual files, the package has a R6 class where each object represents one MicroPEM file, with summary and plot methods including interactivity thanks to **rbokeh**. The package also contains a **Shiny** app for exploration by non-experienced *R* users. The **Shiny** app includes a tab with tuneable alarms, e.g. "Nephelometer slope was not 3" which empowered rapid checks after a day on the field. For later stages of a study after a bunch of files has been collected, the package supports the creation of a measurements and a settings *data.frames* from all files in a directory. We exemplify data cleaning processes, in particular the framework used for the CHAI project, in a vignette, in a transparency effort.

The package is currently available on Github. Since air pollution sensors that would output csvy (csv file with yaml frontmatter) instead of weird csv; and produce ready-to-use data are currently unavailable, **rtimicropem** can be an example of how to use an *R* package as a central place for best practices, thus fostering reproducibility and harmonization of data cleaning across studies. We also hope it can trigger more use of *R* in the fields of epidemiology and exposure science.

Data Analysis Using Hierarchical Generalized Linear Models with R

Youngjo Lee[1], Lars Ronnegard[2] and Maengseok Noh[3]

1. *Seoul National University, Korea*
2. *Dalarna University, Sweden*
3. *Pukyong National University, Korea*

Keywords: Hierarchical likelihood, Hierarchical Generalized Linear Models, Statistical Modelling

Webpages: https://CRAN.R-project.org/package=hglm, https://CRAN.R-project.org/package=dhglm, https://CRAN.R-project.org/package=mdhglm

Since their introduction, hierarchical generalized linear models (HGLMs) have proven useful in various fields by allowing random effects in regression models. Interest in the topic has grown, and various practical analytical tools have been developed. We have summarized developments within the field and, using data examples, show how to analyse various kinds of data using *R*. The work is currently being published as a monograph. It provides a likelihood approach to advanced statistical modelling including generalized linear models with random effects, survival analysis and frailty models, multivariate HGLMs, factor and structural equation models, robust modelling of random effects, models including penalty and variable selection and hypothesis testing. This example-driven book is aimed primarily at researchers and graduate students, who wish to perform data modelling beyond the frequentist framework, and especially for those searching for a bridge between Bayesian and frequentist statistics.

Part III

Posters

Software Development by the Numbers

Adnan Fiaz

There are numerous software development methods out there, each of which focus on creating software as efficient and fast as possible. Waterfall, rapid application development, agile/scrum are a few examples and the evolution of these methods continues. Although some of these methods do involve data this evolution has not yet embraced data. In this talk I will explain and explore the possibilities of using data analysis and analytics to improve the software development process. R and its modelling capabilities are very much at the centre of this exploration. As R becomes more and more popular, R developers will have to work with more established practices. Armed with the techniques from this talk these developers can be a great addition to development teams.

Plot it to Understand it Better: Creating Visualizations in *R* to Support Students in Interpreting Results of Machine Learning Algorithms

Ágnes Salánki[1]

1. R-Ladies Budapest

Keywords: Visualization, Machine learning, Education

Education (and learning) of machine learning algorithms with *R* is quite straightforward nowadays: several good books and great packages support both teachers and students in the process. The introductory algorithms (k-means, decision trees, etc.) are simple, include very few mathematical formula and are easily interpretable on built-in data sets with low dimension number.

Then students go home, pick an interesting data set for their home assignment and try to apply these techniques on their own. At this point, they frequently face obstacles in interpreting the fitted model and getting hints about how they could improve their model.

Since our students use *R* as their primary language for data transformation, exploratory analysis and model fitting (using built-in packages or calling *H2O* algorithms), it was a natural step to map the results of the algorithms into R visualizations they are already familiar with.

Although some really impressive work has been done in the field of visualization of machine learning algorithm results (from the simplest ROC curve to full frameworks, e.g. Noris (2013), Smilkov and Carter (2016), Yee and Chu (2016)), these usually focus on general comparison of techniques like Mülller and Varoquaux (2016) and do not reflect the original decision making process and characteristics of individual algorithms.

The poster presents how standard visualization methods in *R* (like scatterplot matrices, parallel coordinates and tableplots) can be used to support interpretation of the frequently taught machine learning algorithms and summarizes their practical usability from a scalability point of view.

References

Mülller, Andreas, and Gaël Varoquaux. 2016. "Classifier Comparison." http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html.

Noris, Basilio. 2013. "Machine Learning Demos." http://mldemos.b4silio.com/.

Smilkov, Daniel, and Shan Carter. 2016. "TensorFlow Playground." http://playground.tensorflow.org/.

Yee, Stephanie, and Tony Chu. 2016. "A Visual Introduction to Machine Learning." http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

An online translational tool for time to event data

Alberto Alvarez-Iglesias[1] *Amirhossein Jalali*[2] *and John Newell*[1,2]

1. HRB Clinical Research Facility, National University of Ireland Galway
2. National University of Ireland Galway

Keywords: Survival analysis, mean residual life, generalised Pareto distribution, translational statistics

Webpages: https://amir.shinyapps.io/SurvTrans_tool/

Translational Medicine is a concept in the field of biomedical and public health research that conveys the idea that, for any scientific discovery to be fully operational and to reach its goal of improving population health, there has to be efficient ways to transfer the knowledge on the benefits and risks of the newly discovered therapies to the relevant members of society (patients, policy makers, etc.). A similar term, Translational Statistics, can also be used in Applied Statistics where the goal is to develop tools that facilitate the communication of complicated statistical findings to a non-technical audience (Newell et al. 2014; McCabe 2014).

An example of this is found in the analysis of survival data, where one of the challenges is to present meaningful and easily interpretable summaries, avoiding hazard ratios or probability scales that are very often not easily understood. The mean residual life function (MRL), which at any time t is the mean remaining life time given survival up to time t, enables the estimation of very natural summaries, providing answers to the question "how long do I have left" in units of time.

The estimation of the MRL function, however, can be problematic when censoring is due to the termination of the study (type I censoring). In this setting, the use of parametric models or the use of truncated estimates of the survival function will produce undesirable biased estimates of the MRL function. Several hybrid estimators have been proposed to overcome this problem. In this presentation, we focus on one of these estimators that combine existing non-parametric methods with an extreme value tail model (Alvarez-Iglesias et al. 2015). This approach will be demonstrated with a clinical example using an online translational tool that incorporates, along with the estimated MRL, other useful survival plots.

References

Alvarez-Iglesias, Alberto, John Newell, Carl Scarrott, and John Hinde. 2015. "Summarising Censored Survival Data Using the Mean Residual Life Function." *Statistics in Medicine* 34 (11): 1965–76. doi:10.1002/sim.6431.

McCabe, George P. 2014. "Translational Statistics." *Journal of Translational Medicine & Epidemiology* 2 (1): 1022.

Newell, John, Amirhossein Jalali, Alberto Alvarez-Iglesias, Martin O'Donnell, and John Hinde. 2014. "Translational Statistics and Dynamic Nomograms." In *34th Conference on Applied Statistics in Ireland (CASI)*.

Bike sharing usage with Shiny and Leaflet

Alexander Kruse

Keywords: Leaflet, Shiny, Open Data, Spatial Analysis, Maps

My interactive map shows the bike sharing usage of StadtRAD, the bike sharing system in Hamburg – Germany. The data is available on the open data platform from Deutsche Bahn, the public railway company in Germany.

From data processing and spatial analysis to visualization the whole project was done in *R*. I have used the **leaflet** and **shiny** package to display the data interactively. The bikes themselves don't have GPS, so the routes are estimated on a shortest route basis using the awesome CycleStreets API. The biggest challenge has been the aggregation of overlapping routes. I found the `overline` function from the **stplanr** package very helpful. It converts a series of overlaying lines and aggregates their values for overlapping segments.

You can find the whole code from processing to the shiny functions on my GitHub and a write up on my blog.

Use of R at trivago

Alex Dolphin, Michael Frings, Peter Brejcak, Toni Linnenbruegger

Keywords: trivago, poster

Webpages: http://company.trivago.de/

The trivago poster session will address the following topics:

- How R helps the largest hotel metasearch company to make data understandable for our colleagues
- How we detect anomalies
- How we use R with our big data systems
- How we apply machine learning techinques to improve bidding algorithms

Routing Along Multiple Paths

Andreas Petutschnig[1] *and Mark Padgham*[1]

1. Department of Geoinformatics, University of Salzburg

Keywords: Routing, Graph, Network, Probabilistic Routing

Webpages: https://github.com/osm-router/osmprob

When using routing algorithms to model movement on a graph, one is not necessarily interested in obtaining the least-cost, but the most realistic results. The average path between any two points is likely never the shortest, yet there is currently no way to estimate the course or lengths of realistic paths. Instead, movement is often assumed to be singular and repeatable. **osmprob** is an *R* library that provides routing probabilities along all possible paths between any two points. The path of highest probability is the shortest, yet all edges will generally have non-zero probabilities of being traversed. Importantly, the resultant network enables calculation of the probabalistically most likely distance travelled between any two points. The package comes with a built-in **Shiny Leaflet** application to display the results in a matter that is both understandable and visually pleasing.

Impact of biased sampling effort and spatial uncertainty of locations on models of plant invasion patterns in Croatia

Andreja Radović[1], *Stefan Schindler* [2,3], *David Rossiter* [4,5] *and Toni Nikolić*[6]

1. *Czech University of Life Sciences Prague* radovic@fzp.czu.cz
2. *Environment Agency Austria*
3. *University of Vienna*
4. *Cornell University*
5. *University of Twente*
6. *University of Zagreb*

Key words: Biodiversity databases, Balkans, data quality, regression kriging, spatial analysis.

Biological databases are often used in analysing distribution of different taxa but are usually characterised by variable sampling effort and spatial uncertainty of locations. We tested the influence of geographically biased sampling effort and spatial uncertainty of locations, on models of species richness. For this purpose we assessed the pattern of invasive alien plants in Croatia using Flora Croatica Database. The procedure of testing sensitivity of models consisted of tessellating the area into coherent ecological classes (hereinafter Gower classes); ranking quadrants according to sampling effort per class; creating models using varying numbers of quadrants and testing their performances with independent validation points; determining a best fitting model and a threshold of sampling effort, below which data are too unreliable for modelling; simulating spatial uncertainty by adding an adequate random term to each location; and re-running the models by using the simulated locations.

Biased sampling effort and spatial uncertainty of locations had similar effects on model performance regarding the magnitude of the area affected, as in both cases 7% of the quadrants showed statistically significant deviations in invasive alien plant species richness. Our most reliable model, predicted a mean number of 3.2 species for the Alpine region; 5.2 for the Continental; 6.1 for the Mediterranean and 5.3 for the Pannonian region of Croatia. Observational databases can thus be considered a reliable source for ecological models, if their limitations are carefully considered. In order to obtain precise estimates of species richness it is necessary to sample the whole range of ecological conditions.

Comparison of different variable selection strategies to formulate predictive models in medicine

Anita Windhorst[1] and Jörn Pons-Kühnemann[1]

1. Institute of Medical Informatics, Justus Liebig University Giessen, Germany

Keywords: Multivariate analysis, sparse partial least squares analysis, predictive analysis of microarrays, stepwise selection

Physicians are frequently faced with a high number of clinical parameters (e.g. data from the laboratory, radiology, bioinformatics) to aid their decisions, with regard to the best treatment strategies, or to predict the outcome of certain health conditions. These parameters are often highly correlated and connected to a general condition of the patient. In recent years high-throughput methods, such as transcriptome analysis, have gained increased importance in the day to day work of physicians. Several statistical methods have been developed/adapted and implemented in *R* packages to aid the interpretation of this kind of data. Still, even from a statistician point of view the interpretation of the data is not a trivial task. Frequently a high number of transcripts are differentially regulated, which makes it difficult to select suitable predictors for the outcome in question. The identification of robust predictors connected to certain health conditions that can be used in the daily clinical routine, has to precede any application.

Using the example of preterm birth we now want to compare different variable selection strategies adding this process and highlight how R and the existing libraries could be used to achieve this goal. Existing selection strategies will be compared, including sparse partial least squares discriminant analysis (sPLSDA) using the **mixOmics** package, predictive analysis of microarrays (PAM) with **pamr**, and stepwise selection using the Aikaike Criterion (AIC) in order to obtain the best model for the prediction of clinical outcomes.

Shiny as a Medium for Simplifying Disease Rating Scale Computations

Ann Marie Weideman[1], Tan Tran[2], Chris Barbour[1,2], Bibiana Bielekova[1]

1. Neuroimmunological Diseases Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD
2. Department of Mathematical Sciences, Montana State University, Bozeman, MT

Publications in fields related to disease research are growing at an exponential rate. Researchers have turned to statistical analysis and machine learning to develop more sensitive scales for measuring disease progression. However, the underlying message of the research is frequently buried within a script hidden at the tail-end of the Supplementary Material. Moreover, many of the scales involved in disease research require computational power that cannot realistically be achieved with a scientific calculator. Consequently, the end-users (frequently clinicians who are in need of the predictive power of these scales) turn to the most accurate, yet reasonably computable method. As a result, computationally-intensive methods curtail forward movement in disease research, and many novel rating scales are swiftly abandoned for older, more familiar methods. To overcome these challenges in our own research, we developed a user-friendly web-application using the Shiny framework in R (Chang, Cheng, Allarie, Xie, & McPherson, 2017). The purpose of this project was to enable ease-of-access to multiple sclerosis (MS) scales such as Combinatorial MRI Scale (COMRIS) (Kosa et al., 2015), Combinatorial Weight-Adjusted Disability Score (CombiWISE) (Kosa et al., 2016), and the more recently developed Multiple Sclerosis Disease Severity Scale (MS-DSS) (Weideman et al., 2017). Moreover, clinicians can input information regarding the initiation age and efficacy of administered therapy in order to anticipate disease progression at specific time points. We envision that this tool will provide clinicians (and their patients) with a graphical representation of the patient's projected disease course and ultimately improve clinical outcomes. The goal of this presentation is to petition the use of web-applications within the scientific community, particularly for those involved in research of neurological diseases. Shiny is extremely viable for the laboratory or clinical setting, as it does not require extensive knowledge of Javascript or markup languages. Despite these features, however, we encountered several handicaps throughout the development process that we believe to be worth sharing. In addition, we discuss the resources that we found to be most helpful during the web-design phase, from planning the sitemap to debugging and deploying the application. We are optimistic that laboratories within the National Institutes of Health (NIH) and other research communities will adopt similar methods for communicating their research to those involved in clinical care.

References

Chang, W., Cheng, J., Allarie, J., Xie, Y., & McPherson, J. (2017). shiny: Web Application Framework for R. (Version R package version 1.0.0). Retrieved from https://CRAN.R-project.org/package=shiny.

Kosa, P., Ghazali, D., Tanigawa, M., Barbour, C., Cortese, I., Kelley, W., et al (2016). Development of a Sensitive Outcome for Economical Drug Screening for Progressive Multiple Sclerosis Treatment. Front Neurol, 7, 131. doi:10.3389/fneur.2016.00131

Kosa, P., Komori, M., Waters, R., Wu, T., Cortese, I., Ohayon, J., et al (2015). Novel composite MRI scale correlates highly with disability in multiple sclerosis patients. Multiple sclerosis and related disorders, 4(6), 526-535. doi:10.1016/j.msard.2015.08.009

Weideman, A. M., Tapia-Maltos, M. A., Barbour, C., Tran, T., Kosa, P., Komori, M., et al (2017). New multiple sclerosis disease severity scale predicts future accumulation of disability. Unpublished Manuscript.

Introduction githubinstall package designed in JapanR user group

Atsushi Hayakawa[1,2]

1. Recruit Communications Co., Ltd.
2. HOXO-M Inc.

Keywords: github, install, community, Japan, Books

Webpages: https://cran.r-project.org/web/packages/githubinstall/index.html, http://rpkg.gepuro.net/

I would like to introduce you githubinstall packages desined in JapanR user group.

Githubinstall is a package that can help when you install the package from github easily.When you install the package with devtools, you mush specify a user and package name in github. However, by using the githubinstall, you can install the package without a username using crawler gets the parts of username and package name every day. Frequently, we do not remember the developer's name, even in package name was remembered. Therefore, we must look the username each time to install the package to the new environment using devtools. By using this githubinstall package, you can save the time and effort for checking them. If there are same package name in github, you can select any one. The instructions are as follows.

```
library("githubinstall")
githubinstall("dplyr")
```

Japan.R is the R language user group in Japan. In Japan, R user meeting groups exist in some local community, and once a year, Japan.R is held as a general meeting of Japanese R user group. I am the organizer of this Japan.R. We have a website http://japanr.net written in Japanese. In Japan.R 2016, 244 participants and 25 presenters participated, and there came from the more than 13 regions of the local community. For example, there are Tokyo.R that held once a month, meetup event in English that is held when visitor come from abroad, etc. in a part of the local community, and Japan.R is co-hosted with these members In addition, books in Japanese about the R language have been enhanced. For example, Reverse Handbook in R language is 改訂3版 R言語逆引きハンドブック, Book for advanced user is パーフェクトR, and Advanced R written in Hadley Wickham was translated to Japnese is R言語徹底解説. Many writers belonging to the R community have contributed to these books.

Quantifying radiation health effects in the Life Span Study of atomic bomb survivors

Benjamin French[1], Munechika Misumi[1], Kyoji Furukawa[1] and John B Cologne[1]

1. Department of Statistics, Radiation Effects Research Foundation

Keywords: cancer incidence, generalized non-linear models, regression

Webpage: http://www.rerf.jp/library/dl_e/index.html

Understanding the health effects of radiation exposure is important for establishing recommendations for radiation protection, including limits on occupational exposure to radiation and guidelines for diagnostic and therapeutic uses of radiation. The Life Span Study includes residents of Hiroshima and Nagasaki, Japan, who were within 10 km of the hypocenter at the time of the atomic bombings on August 6, 1945, in Hiroshima and August 9, 1945, in Nagasaki. For these survivors, DS02R1 radiation dose estimates, calculated using the DS02 dosimetry system, are based on the survivor's reported location, the amount and type of shielding between the survivor and the blast, and the orientation of the survivor relative to the direction of the blast. The Life Span Study also includes a sample of Hiroshima and Nagasaki residents who were 'not in city' at the time of the bombings. These residents are thought to be representative of the cities' general population, but without exposure to radiation from the bombs. We analyzed the association between radiation exposure and incidence of first primary solid cancer among 80,205 survivors whose DS02R1 radiation dose could be calculated and 25,239 not-in-city residents. All analyses were based on a highly stratified table of case counts and accrued person-years. Using the *R* package **gnm**, we fit piecewise constant hazard models to quantify heterogeneity in cancer risk according to sex, age, city, and location (e.g, 'not in city'), and to estimate the excess relative risk of cancer associated with radiation dose. Effect modification of the excess relative risk by sex and age was considered. A regression model with internal standardization provided a sex-averaged excess relative risk of 0.501, 95% CI: (0.405, 0.601) per Gray of weighted absorbed colon dose, as well as strong evidence of a curvilinear dose response among males (P=0.008). Our analysis illustrates the flexibility of **gnm**, along with user-specified functions for the model terms, to formulate non-linear regression models.

Exploring TIMSS at the Question-Level With R

Benjamin Ortiz Ulloa

Keywords: TIMSS, tidyverse, data wrangling, machine learning, visualization

Websites: http://rms.iea-dpc.org/, https://beemyfriend.github.io/Articles/Visualizing_2011_TIMSS.html

Abstract: The Trends in International Math and Science Study (TIMSS) is a series of international assessments given to 4th grade and 8th grade students around the world. It is administered by both the International Association for the Evaluation of Educational Achievement (IEA) and Boston College (BC) who first conducted the assessments in 1995 and re-administered them every 4 years after that (1999, 2003, 2007, 2011, and 2015). The TIMSS dataset contains information on how every participating student answered each question on the test. The dataset also provides supplemental information on the individual student, the student's teacher, student's school, and the student's country. While the TIMSS dataset is provided in only *SPSS* and *SAS* formats, suggesting that most analysis is done in those languages, I will use *R* to explore the TIMSS dataset. I will show how different tools in the *R* ecosystem can be used to tidy the TIMSS dataset as well as create and explore new insights in an already deep dataset. I will take a particularly deep look into how individual students and countries performed on each question. Once the dataset is cleaned, I will create different models in an attempt to predict how a student will perform on a particular question on the TIMSS tests. The process of tidying and modeling the TIMSS dataset will show how quickly and easily such tasks can be done in *R*.

doAzureParallel - Using Azure as Your Parallel Backend for high performance work

Brian Hoang brhoan@microsoft.com

Github Pages: https://github.com/Azure/rAzureBatch https://github.com/Azure/doAzureParallel

Description: Using hundreds or thousands of cores is common for High Performance Computing (HPC) workloads for financial modeling. Monte Carlo simulations for regulatory risk calculations, modeling financial instruments, regression testing or managing portfolio are typical, but the cost and complexity to scale up and scale out R code is difficult to access.

The doAzureParallel package is a parallel backend for the widely popular foreach package that lets users execute multiple processes across Azure VMs. In just a few lines of code, doAzureParallel helps users to create and manage their cluster in Azure, and scale their work to up to hundreds or thousands of cores. doAzureParallel is ideal for running embarrassingly parallel work such as parametric sweeps or monte carlo simulations, making it a great fit for many financial modelling algorithms.

By using Azure as a parallel backend, users can take advantage of Azure's high performance machines and scale out their workloads as much as they need. Using Azure allows users to use the cloud's elasticity. The doAzureParallel package lets users take advantage of Azure's auto-scaling capabilities, increasing and decreasing the cluster size to fit the user's workload. Users only pay for the compute cycles they consume.

Rtraildb: R interface to TrailDB

Bryan Galvin

AdRoll

Keywords: big data, high performance computing, operationalizing R, etl

Webpages: http://traildb.io https://github.com/traildb

Abstract: Rtraildb is an R package that provides a wrapper for TrailDB, an efficient tool for handling event data. AdRoll created TrailDB to power its analytic workload of web traffic data and leverages it to store and query over a quadrillion events per year. Put simply, TrailDB is a read-only file that is designed to be a building block for systems that need to process a large number of discrete events organized by a primary key. Immutability allows for deeper compression, scalability and architectural decisions that aren't feasible with existing databases.

TrailDB Feature Highlights:

- High performance: Process millions of events per second on a single core
- High compression ratio: Comparable to Gzip
- Scalable: Capacity of your server is the limit
- Immutable: Easy to handle with distributed architectures

With Rtraildb, R users will be able to perform exploratory data analysis on highly compressed data without having to bring everything into memory. In addition, projects built on top of TrailDB such as reel and trck allow for even more extensive querying that would be hard to express using dplyr, data.table or even SQL syntax (for example: *count of event B following event A within 5 seconds by group*). This presentation will introduce Rtraildb and cover use cases where it might be a good option in place of a traditional relational or timeseries database.

Using R for an efficient and standard analysis of Design of Experiments.

C.Mokdad[1], N.Antille[1], M.Lepage[1], M.Perrot[1], N.Pineau[1] and H.Vlaeminck[1]

1. Nestlé Research Center, Lausanne, Switzerland

Keywords: Design of experiments, linear modeling, least significant difference, principal component analysis, shiny

Design of Experiments (DOE) is a useful strategy to guide R&D specialists through the planning and execution of experiments (trials) and helps them to take data-driven decisions. In food processing, one is generally seeking to assess a set of operating parameters in different processes, which takes time and resources. Instead of a trial and error approach, the DOE helps to get insightful results about the global impact of parameters on the final product. This approach relies on well-established mathematical concepts which enable to maximize information from a limited number of trials.

Data scientists at Nestl? R&D have applied this methodology for many years, and have found that the R environment is particularly suited for the efficient analysis of such DoE. R is a flexible platform, providing high quality outputs, which can also use in industry to automatize our workflows. In the context of DoE analysis, the following one was developed. First, a linear mixed model is used on the raw sensory data to allow us to find significant differences between products (using Fisher's Least Significant Difference as a post-hoc test)*. Then, a model (usually linear) on the mean sensory data enables to assess the relative importance of the factors of the DOE, and to quantify the effect size of each factor on the sensory. In addition, the sensory data is represented in a 2D space (Principal Components Analysis) that represents the sensory space generated by the products coming from the DOE.

R enables to use standard outputs (plots and models) and to efficiently share and reuse this workflow among the data scientists. Moreover, R also allows to develop shiny tools which provide a user-friendly interface with stakeholders. An internal shiny tools platform was developed to access these tools.

A case-study, relating sensory characteristics to process parameters of products, will illustrate how different R functions and packages are used in the workflow previously presented.

Easy and Stable Deployment of *R*-based Software with *github*, *devtools*, and *docker*

Daeyoung Kim, Jinsung Kim, Seonjeong Lee, and Hyoseop Lee

Encored Technologies, Inc. Seoul, Korea

Keywords: NILM, devtools, deployment, dependency

Webpages: http://enertalk.us

NILM (Non-Intrusive Load Monitoring) is a technology that breaks down a household's aggregate electricity consumption into individual appliance's electricity consumption. It provides useful information for energy savings and life pattern analysis while avoiding the cost and trouble of installing numerous measurement devices to every individual appliances. Currently, we provide NILM service to +20,000 homes in Korea with *R*-based software on several dozens of servers.

The NILM service, like any other machine learning based services, operates in perpetual development mode, in which algorithms are continuously improved and make them available to users. Consequently, rapid and stable deployment is one of chief challenges that algorithm developers face.

In this poster we propose efficient deployment of the *R* software, focusing on *github* based code with devtools for package installation. The main challenges are 1)resolving dependency for private github code base with *devtools*, and 2)prevention of conflict of dependency between different products. The problem 1) is described as a series of issues like https://github.com/hadley/devtools/issues/1382, and solutions are followed by like https://github.com/hadley/devtools/pull/1382. *Docker* based virtualization is proposed to a solution of 2), which also enables software update by replacement of images. Overall, we propose a simple and stable way for rapid deployment of *R* software with the combination of github, devtools, and docker.

A Mixture of Multivariate Regressions Package with Applications for Chronic Kidney Disease

UseR! Conference, Brussels, Belgium

David B. King [1] **and Jefferson Davis** [2]
[1] Clinical Assistant Professor, Epidemiology And Biostatistics, Indiana University
[2] Research Analytics, Indiana University

Chronic kidney disease (CKD) is a physiological condition in which the kidneys ability to eliminate wastes from the blood gradually declines to critical levels. At present the disease is most often diagnosed on the basis of a single biomarker: either the glomerular filtration rate (GFR) or an estimate of the glomerular filtration rate (eGFR). However, there are many other biomarkers associated with CKD such as systolic blood pressure, diastolic blood pressure, and fasting plasma glucose. Looking at the values of these biomarkers in populations with and without CKD we see multidimensional ellipses with distinct means and covariances. This observation leads to a model in which for the two populations Π_1 and Π_2 of individuals with and without CKD and the multivariate measure of biomarkers $\mathbf{Y}$, $[\mathbf{Y}|\Pi_i] \sim N(\mu_i, \Sigma_i)$ for $i = 1, 2$. In a diagnostic situation we care about which population an individual is in given $\mathbf{Y}$. So we apply Gaussian mixture modeling where we compute $P(\Pi_i|\mathbf{Y})$ for $i = 1, 2$ and classify individuals according to a simple decision rule, such as classifying an observation as Π_1 if $P(\Pi_1|\mathbf{Y}) > 0.5$.

Furthermore we often encounter covariates in medical settings BXsuch as age, sex, or BMIthat we wish to factor in the model. In this situation, we create a Mixture of Multivariate Regressions (MMR) and then use this mixture model to obtain $P(\Pi_1|\mathbf{Y}, \mathbf{X})$ and $P(\Pi_2|\mathbf{Y}, \mathbf{X})$.

We discuss details associated with the model and details of the implementation in R and how we extend the package *mixtools*. We will illustrate our algorithm with an analysis of chronic kidney disease diagnosis and show that this algorithm has the potential to sharpen diagnoses within a wide medical community.

Ranking influential communities in networks

David Selby[1] *and David Firth*[1]

1. Department of Statistics, University of Warwick, UK

Keywords: PageRank, Bradley–Terry model, networks, ranking, bibliometrics

Webpages: http://warwick.ac.uk/dselby

Which scientific fields export the most intellectual influence, through recent research, to other fields? Citation behaviour varies greatly across disciplines, making inter-field comparisons difficult. Recent approaches to measuring influence in networks, based on the PageRank algorithm, take the source of citations (or recommendations or links) into account.

By aggregating all publications in each Web of Science field into "super-journals", we model the exchange of citations between fields. We then fit a Bradley–Terry paired comparisons model—which can be shown to be equivalent to scaled PageRank—to measure the relative influence of academic communities. Uncertainty intervals are provided to quantify the robustness of the ranking. All analyses were performed in R.

Which field is top of the table? Visit the poster to find out!

References

Bergstrom, Carl T, Jevin D West, and Marc A Wiseman. 2008. "The Eigenfactor Metrics." *The Journal of Neuroscience* 28 (45). Society for Neuroscience: 11433–4.

Varin, Cristiano, Manuela Cattelan, and David Firth. 2016. "Statistical Modelling of Citation Exchange Between Statistics Journals." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179 (1). John Wiley & Sons, Ltd: 1–63.

Integration of an *R* forecasting engine in the *.NET* back-end of *c-Quilibrium*'s cash supply chain optimization solution

Davy Sannen[1] *and Andries Van Humbeeck*[1]

1. *AE - architects for business & ICT*

Keywords: Operationalization, Integration of *R* and *.NET*, Parallelization

Webpages: http://analytics.ae.be

A data analysis project does not always end after the analysis phase. Typically your newly developed *R* algorithms must be integrated into an operational production system before they can realize their added value. This presents us with some additional challenges to be tackled.

In this talk we will discuss the integration of an *R* forecasting engine in the *.NET* back-end of *c-Quilibrium*'s cash supply chain optimization solution. Several topics will be addressed, including the communication between *R* and *.NET*, parallelization of the *R* algorithms, logging and exception handling. We will discuss the choices we made, the *R* packages we used and the lessons learnt by applying them in an operational environment.

Modeling cage effects in microbial community in Type 1 Diabetes (T1D) mouse model using mixed effect models

Dea Rynanda Putri[1,2], Ziv Skhedy[2], Olivier Thas[3], Nolen Joy Perualila[2], Thierry van Effelterre[1], Xuesong Zhang[4], Marcus Rauch[5], Martin J. Blaser[4], Luc Bijnens[1]

1. *Janssen Research and Development, Beerse, Belgium*
2. *Hasselt University, Diepenbeek, Belgium*
3. *Ghent University, Ghent, Belgium*
4. *New York University Langone Medical Center, New York, USA*
5. *Janssen Prevention Center, London, UK*

Keywords: cage, time, microbiome, mouse model, mixed effect models

Murine models are a crucial component of gut microbiome research. Unfortunately, a multitude of genetic backgrounds as well as experimental setup, together with inter-individual variation complicate comparison between studies and a global understanding of the mice microbiota. McCafferty et al.(2013) suggested that maternal transmission and cage effects are important confounding factors in microbiome studies. To assess the cage microenvironment effect on the mouse gut microbiome, we collected fecal samples from 178 same breed mice for 49 days (bi-monthly fecal samples collection). The mice were grouped into control and antibiotics group. Two modeling approaches were used. In the first, the cage effect was assessed on the level of alpha diversity (richness). A mixed effect model, in which the cage effect is included as a random effect, was used to detect potential correlation between animals sharing the same cage. Even under these controlled conditions, strong cage effect on alpha diversity was found after correcting for the treatment effect. The second modeling approach focused on specific models by OTU. In particular, we investigated whether cage also has a significant effect on the OTU appearance between two consecutive time points. A negative binomial linear mixed effect models was used in which cage effects were assumed to be random effects. A strong cage effect was found in 57 out of 348 OTUs after correcting for multiple tests. Our finding suggests that even in the controlled population of same breed laboratory animals, the gut microbiota composition may not be as uniform as previously thought. We believe that more contributing factors may yet have to be identified to explain additional components of variation in the composition of the microbiota within populations and individuals over time. These findings might have implications for the design and interpretation of experiments when it involves laboratory animals, even in a controlled environment.

References

McCafferty, J., M. Muhlbauer, R.Z. Gharaibeh, J.C. Arthur, E. Perez-Chanona, W. Sha, C. Jobin, and A.A. Fodor. 2013. "Stochastic Changes over Time and Not Founder Effects Drive Cage Effects in Microbial Community Assembly in a Mouse Model" 7. The ISME Journal: 2116–25.

Using **R-INLA** to understand institutional moderators of drought

Emily Burchfield[1] *and Katherine Nelson*[1]
1. Department of Civil and Environmental Engineering, Vanderbilt University

Keywords: **R-INLA**, Drought, Spatio-temporal, Bayesian

Webpages: https://ekburchfield.wordpress.com/,
https://github.com/eburchfield/CA_drought

California's Central Valley region has been called the "bread-basket" of the United States. The region is home to one of the most productive agricultural systems on the planet. Such high levels of agricultural productivity require large amounts of fresh water for irrigation. However, the long-term availability of water required to sustain high levels of agricultural production is being called into question following the latest drought in California. In this presentation, we use Bayesian multilevel spatiotemporal modeling techniques to examine the influence of the structure of surface water rights in the Central Valley on agricultural production during the recent drought.

Annual data for years 2007-2014 of the recent drought were obtained for the entire Central Valley with outcome, control, and predictor variables available at one of two different spatial scales: field-level (1-km pixels) or watershed level (USGS HUC-12 designation). To capture field-level production dynamics, we computed an index of total vegetative production (TVP) using remotely sensed metrics of vegetation health from the MOD13A1 MODIS dataset. Point data identifying the location of surface water right points of diversion (PODs) and the legal status of each POD were downloaded from the CA SWRCB electronic water rights information management system (eWRIMS). Other key predictors include cumulative drought stress as measured by the Standardized Precipitation Index (SPI), land use and agricultural diversity within a watershed, and a novel gridded groundwater dataset constructed using the `spacetime` package (Bivand, Pebesma, and Gomez-Rubio 2013).

To examine the effects of the structure of water rights on agricultural productivity during times of water scarcity the observed TVP was fit to a multi-level model with water right-SPI interactions, which can be expressed generally as:

$$y_{ijk} = \beta_{0jk} + \beta_{10k}SPI + \beta_{20k}X + \beta_{30k}X * SPI + \beta_{4jk}C + s_{00k} + e_{ijk} \qquad (1)$$

where, β_{0jk} is an intercept term, β_{1k} represents the linear effect of cumulative meteorological drought stress (SPI) on TVP, β_{20k} is a vector of coefficients that describe the effects of water rights on TVP at the watershed level, X is a vector of water rights predictors (Percent Riparian, Percent Pre-1914, and Percent Appropriative), β_{30k} is a vector of coefficients that describe the effect of interactions between water rights predictors and SPI, β_{4jk} is a vector of coefficients for controlling variables, C is a vector of controlling variables (year, land use category, water rights density, agricultural diversity, and annual groundwater elevation change), s_{00k} is a watershed level spatial effect, and e_{ijk} is a random effect accounting for within field variability. In order to account for complex spatial effects modeling was performed using the R package `R-INLA` (Blangiardo and Cameletti 2015). Spatial effects at the watershed level were modeled using an intrinsic conditional autoregressive (iCAR) model coupled with an exchangeable (iid) random effect, also known as a Besag-York-Mollié (BYM) model.

References

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. http://www.asdar-book.org/.

Blangiardo, Marta, and Michela Cameletti. 2015. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Wiley. https://sites.google.com/a/r-inla.org/stbook/.

The Analysis of the Determinants of Exchange Rate via Conditional and Partial Granger Causality Test

Erdogan CEVHER[1]

1. The Ministry of Science, Industry and Technology of Turkey

Keywords: Exchange Rate, Conditional Granger Causality, Partial Granger Causality, causfinder.

Webpages:
https://zenodo.org/record/322576,
https://zenodo.org/record/35599,
https://www.academia.edu/9890764/causfinder_An_R_package_for_Systemwise_Analysis_of_Conditional_and_Partial_Granger_Causalities

In this paper, the relationships among the macroeconomic variables (exchange rate, deposit rate, gold prices, BIST100 Istanbul stock market index) in the 2010.08–2015.12 period were investigated by conditional and partial Granger causality tests. First, the theoretical information about conditional and partial G-causality tests is given. Then, these causality tests were applied to the data using the **causfinder** package in *R* software program. The results of the conditional and partial G-causality tests showed significant G-causalities.

References

Guo, S., Seth A.K. and Kendrick K.M. et al. (2008). Partial Granger causality: eliminating exogenous inputs and latent variables. Journal of Neuroscience Methods, 172, 79-93.
http://www.sciencedirect.com/science/article/pii/S0165027008002379

Youssofzadeh, V., Prasad G., Naeem M. et al. (2013). Partial Granger causality analysis for brain connectivity based on event related potentials. 6th INCF Congress of Neuroinformatics, Karolinska Institutet, Stockholm, Sweden. http://www.frontiersin.org/10.3389/conf.fninf.2013.09.00114/event_abstract

How students learn Statistics: from tracing student's activity in R Commander to the visualization of their work through a Shiny app

Martori F.[1], Cuadros J.[1], Calvo M.[2], Miñarro A.[2], Serrano V.[1] and Gorina V.[1]

1. ASISTEMBE, IQS - Univ. Ramon Llull
2. Department of Genetics, Microbiology and Statistics. University of Barcelona

Keywords: Teaching Statistics, Shiny, Learning Analytics

Webpages: http://asistembe2.iqs.edu/rcmdrtr/rcmdrbd_demo/

R Commander(Fox 2005) is the most well-known GUI for teaching and learning statistics as it is used all over the world in many Statistics courses. **R Commander** allows the use of *R* without compromising the learning process as command line interface to *R* can be an obstacle to many students (Fox 2005). We have tweaked **R Commander** so that an activity log of the students' work is obtained. This version of **R Commander** is used in class to trace the students' work in a planned activity. In order to analyze the work done by students, the instructor may design some milestones. These milestones can be understood as steps in the solving process that identify important achievements of the activity resolution.

The **shiny** (Chang et al. 2017) app we are presenting provides a user-friendly interface to quickly visually the work done by students and identify those who may require special attention from the instructor. This information consists of the most common functions used to solve the activity, each student's list of actions and when they occurred and charts displaying the amount of time spent working, the amount of actions done and how they are related.

If a set of milestones is specified, the instructor will also obtain information about the proportion of students who reached each milestone, how do different students' performances cluster and a chart displaying the relationship between student's workload, the time spent and a grade estimate.

The use of this tool should allow Statistics instructors to provide better and personalized feedback to students about their learning process.

References

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2017. *Shiny: Web Application Framework for R.* https://CRAN.R-project.org/package=shiny.

Fox, J. 2005. "The R Commander: A Basic Statistics Graphical User Interface to R." *Journal of Statistical Software* 14 (9). doi:10.18637/jss.v014.i09.

R in support of civil society strengthening in the gender-based violence sector in South Africa

Frans van Dunné[1], Aidan Connolly[2], Elizabeth Bosha[2], Ria Schoeman[3] and Craig Carty[2,4]
1. ixpantia
2. The Relevance Network
3. Foundation for Professional Development
4. University of Oxford, Department of Social Policy and Intervention

Keywords: Social Development, Policy and decision making, custom graphics, Shiny, ggplot2

Webpages: https://wwww.ixpantia.com, https://therelevancenetwork.com, http://www.foundation.co.za

Data-driven analyses play an increasingly critical role in social development worldwide. Evidence-generation provides insights into both gaps and best practices, which can translate into improved programming and policy decisions. The reliability and validity of evidence, however, hinges upon numerous underlying assumptions. Perhaps the most challenging of these is the numerical and statistical literacy of those who seek benefit from rigorous social science investigations, namely stakeholders from the public and private sectors. Leveraging the potential of rapid data analytics, coupled with the visualization of findings to make them "accessible" to lay persons, is key to countering the realities of diminished pools of social investment funds.

A collaboration between data scientists, implementing partners, and civil society organizations (CSOs) in the gender-based violence (GBV) sector was launched in 2016 as part of an on-going landscape analysis in South Africa. The goal of the project was to measure the functionality of GBV CSOs across key performance areas: governance, partnership building, advocacy, monitoring & evaluation, and internal capacity. To collect data that would inform robust analyses, the team constructed a web-based questionnaire using modified tools from the NGO sector tailored for GBV service providers (FANIKISHA, NGO Scorecard, and the Independent Code for Non Profit Good Governance). Responses to the online questionnaire (n=122 variables, n=24 GBV CSOs nationwide) were distilled into 10 organizational dimensions and weighted appropriately to support predictions of "sustainability". Using the resulting scores from the focal practice areas within the data set, the team developed a model to translate relatively complex data into visualizations and reports; first, for those capable of performing advanced statistical inquiry, and second, for those lacking numerical literacy.

The evaluation of data accounted for multivariate analyses, the desire to generate custom graphics in ggplot2, and the pragmatic need to tell "stories" using visually-driven narratives informed by experts. The synergies between these three fundamental considerations resulted in an accessible, visual mapping of individual GBV CSO capacity, as well as Principal Component Analyses that generate distance maps to align like-skilled organizations. The purpose of the latter is to support matching well-capacitated GBV CSOs with those in need of strengthening for the purposes of constructing a mentorship/mentee model nationwide. In addition, the layered capacity of the toolset built in R, coupled with the visualizations owing to a Shiny app, supports multiple views of the data from within and across the participating GBV CSOs. This is critical to policymakers and donors alike, who require access to nuanced data interpretations to inform decision-making.

The toolset available in R for these kinds of top-level capacity interrogations is excellent, as it allows for rapid iterations and translations to other practice areas while still offering enough granularity to make precise customizations. Any improvement in the support given to these CSOs owing to the visual interpretations of the data can have a substantial beneficial impact, particularly in settings like South Africa where GBV rates are persistently among the highest in the world (Patience and Nondumiso 2015).

References

Patience, Mpani, and Nsibande Nondumiso. 2015. "Understanding Gender Policy and Gender- Based Violence in South Africa, a Literature Review for Soul City: Institute for Health and Development Communication." Braamfontein, South Africa: Tshwaranang Legal Advocacy Centre.

Using *R* in transdisciplinary approches for visualiazing and analysing people's perceptions, knowledge and practices in complex social-ecological systems

Frédéric M. Vanwindekens[1], Didier Stilmant[1] and Philippe V. Baret[2]

1. Walloon agricultural research centre (CRA-W) – Agriculture and natural environment Department (D3) – Farming Systems, Territory and Information Technologies Unit (U11)
2. Earth and Life Institute, ELIA, University of Louvain (UCL), Louvain-la-Neuve, Belgium

Keywords: qualitative survey, semi-quantitative approach, cognitive mapping, transdisciplinarity, visualization

Webpages: https://CRAN.R-project.org/package=mypkg

Social-ecological systems are complex due to uncertainty related to their nature and their functions. In these systems, decision-making processes and practices of managers are, at least partly, value-laden and subjective, influenced by their world-views and their own knowledge. People's perception, knowledge and practices are central in building their adaptive capacity but are seldom taken into account by traditional decision-making approaches in modelling social-ecological systems management.

We developed an original systemic approach that aims to visualize and analyse perceptions, knowledge and practices of people in social-ecological systems. The method approach is called CMASOP (Cognitive Mapping approach for Analysing Systems Of Practices) and can be used in two interrelated ways: (i) a descriptive and exploratory one and (ii) a comparative and typological one.

The whole approach has been carried out using *R* and we are currently working on the development of a version of CMASOP that could be shared among the scientific community as an *R* package.

As a case study, we applied the method to the study of grassland management in livestock farming systems of Belgium. We showed the relevance of using semi-qualitative modelling tools, like cognitive mapping, for describing and comparing these kind of complex systems.

The main steps of the descriptive utilization of CMASOP are: (1) carrying out of a qualitative survey and transcription of interviews, (2) coding of interviews, using the *R* Qualitative Data Analysis (**RQDA**) tools, (3) Individual Cogntive Mapping based on the list of relations identified in the interviews and (4) Social Cognitive Mapping based on the aggregation of individual maps. Cognitive maps have been produce using *R* packages **network** and **Rgraphviz**. Steps 3' (clustering or categorizing individual maps) and 5 (comparative analysis) are added for using CMASOP as a comparative and typological tool based on people's perception, knowledge or practices. For this comparative utilization, CMASOP require the following *R* packages : **cluster**, **ade4**.

Extensive descriptions of these steps and their application to the case study can be found in Vanwindekens et al. (2013) and Vanwindekens et al. (2014). Our approach has also been applied for studying practices in other social-ecological systems : farmers and biodiversity in France, winegrowers in Italy, beekeepers in France.

References

Vanwindekens, F.M., Stilmant, D. & Baret, P.V. (2013). Development of a broadened cognitive mapping approach for analysing systems of practices in social-ecological systems. Ecological Modelling 250 : 352-362.

Vanwindekens, F.M., Baret, P.V. & Stilmant, D. (2014). A new approach for comparing and categorizing farmers' systems of practice based on cognitive mapping and graph theory indicators. Ecological Modelling 274 : 1-11.

Exploring Europe's comparative advantage in low-carbon technology using R

Georg Zachmann[1] *and Robert Kalcik*[1]

1. Bruegel, Brussels

Keywords: Low-carbon tech, patents, exports, innovation

Webpages: https://bookdown.org/robertkck/ecf_draft/businessmodel.html

Given global decarbonisation concerns, the wide array of low-carbon technologies offers significant growth potential. Some EU countries have already been able to develop a comparative advantage in wind turbines and electric vehicles, though the EU is less effective at exporting solar panels and batteries. Based on patenting activity, we identify potential to further specialise in all of these four low-carbon technologies - maybe not for entire countries but for some regions. A regional overview is presented because it can help in targeting public investment (eg in infrastructure, research and education) to enable development in the most promising sectors and regions. The project leverages *R* throughout the research pipeline from querying the large scale international patent and trade databases, to analysing bipartite networks of countries and patents/exports, and communicating the results with Shiny, Plotly and Bookdown.

Central Bank Communications: Information extraction and Semantic Analysis

Giuseppe Bruno
Bank of Italy, Economics and Statistics Directorate.

Keywords: Text Mining, Semantic Analysis, Pointwise Mutual Information, Web search

Central Banks, among other tasks, provide a relevant amount of information for Institutions and market operators. Indeed Central Banks employ a multiplicity of communication channels to drive market expectations. We deem quite relevant to provide a quantitative evaluation of the impact of these reports in increasing the central bank transparency with the goal of enhancing the effectiveness of its institutional action. In this poster we take the challenge of adopting and experimenting a methodology for quantifying the linguistic content of some official documents of the Bank of Italy. In particular we take into consideration the Financial Stability report (henceforth FSR) which is a young publication whose first issue came out on 2010. Here the goal is twofold: on one hand we provide a transparent numerical framework to consider sub-unit of an official Central Bank report. Moreover it is proposed an analytical tool to gauge the impact of an official document on the public. In the context of the official reports released by the Bank of Italy, we show how this framework can be employed to characterize and extract their information content. Once the documents are decomposed into their constituent sentences, we estimate some figure for the readability and the formality of our reports. We build a small corpus of all the available issues of the FSR and we show how to convert its documents into a vector space model by employing familiar concepts of linear algebra such as eigenvalues and eigenvectors. Among these vector space models, we consider here the Latent Semantic Analysis (henceforth LSA) model. Within the linguistic semantic framework words, statements, chapters and whole documents are represented as high dimensional vectors in the same space. The main advantage of this common representation is the possibility to take advantage of the natural metric induced by the vector space for analyzing semantic relationships. Therefore in the LSA model we can: 1. compute semantic similarity measures between words/documents by exploiting the statistical redundancies in text; 2. compute word neighborhood which are set of words/documents sharing semantic concepts (synonimity); 3. compute text coherence and summary of given documents. We make one step further by taking inspiration from previous works on similar topics [3, 1, 2]. We evaluate the public perceptions associated with every sentences of the FSR and some couples of antithetic keywords relevant for the financial stability issues. The semantic orientation is evaluated employing the web-hit count for estimating the Pointwise Mutual Information. Two of the most relevant issues in measuring sentiment through the web are: the sampling variability and the statistical representativity of the sample. Here we address the sampling variability by comparing the effectiveness of three web search engines: Bing, Google and Yahoo. These three search engines reached about 90% of the market share in 2015. We found some differences between the results provided by Google on one side and those achieved with Bing and Yahoo on the other. For our empirical application we have employed many packages for text mining and sentiment analysis available in R. The main results obtained from this analytical framework look quite promising for extracting information content from the publications. The development of web oriented tools for monitoring and extracting sentiment orientation fosters a wider adoption of text mining and semantic techniques aimed at improving the statistical accuracy required to assume well informed decisions.

References

[1] Carvahlo, C., C. Cordeiro, and J. Vargas (2013). Just words? A quantitative analysis of the communication of the central bank of Brazil. Revista Breasileira de Economia 67(4), 443–455.

[2] Kawamura, K., Y. Kobashi, M. Shizume, and K. Ueda (2016). Strategic central bank communication: Discourse and game-theoretic analyses of the bank of Japan's monthly report. JSPS Working Paper Series (80), 1–34.

[3] Lucca, D. O. and F. Trebbi (2011). Measuring central bank communication: an automated approach with applications to FOMC statements. NBER working paper (15367), 1–37.

On the use of R for building a responsible data science workflow in the retail industry

Hinda Haned

Ahold Delhaize, HR analytics

Keywords: human resources, machine learning, reproducibility, transparency, interpretability

The use of machine learning to answer business questions has become a competitive advantage in many industries; in particular, when it comes to leveraging data to answer human resources (HR) related questions. For example: how can associates' well-being at work be assessed and improved? What is causing absenteeism or turnover? In this context, applying machine learning to HR data raises important questions on the accountability of data scientists:

- How can we balance responsible data science practices in applying machine learning algorithms and (pressure) of utility for business?
- Are there sensitive data sources that should not be collected and used? Where can we (or shall we) draw the line?
- How do we validate our results in a reproducible way?
- How can we communicate results to businesses (when HR data are used) in a transparent and interpretable way?

In this talk, we will discuss how *R* was essential in solving these questions in the context of an HR analytics team of an international retailer. In particular, we will show how relying on *R* **Markdown**, **Shiny** and **flexdashboard**, enabled us to swiftly and concretely develop a workflow where data scientists can take advantage of the power of machine learning tools in *R* packages while safeguarding transparency, reproducibility and interpretability of the results.

We illustrate our workflow with a project involving the analysis of opinion survey data conducted on more than 200,000 associates. Using *R*-powered tools, we processed the survey results, validated its construct and used NLP tools to summarize the text data (LDA, word2vec, sentiment analysis). We further discuss how we selected and presented the results to stakeholders. We will conclude by sharing briefly how the outcome of this project impacted our company's way of working.

CrispRVariants: visualising gene editing experiments with R and shiny

Lindsay H[1,2], Burger A[1], Biyong B, Felker A[1], Hess C[1], Zaugg J, Chiavacci E[1], Anders C[1], Jinek M[1], Mosimann C[1] and Robinson MD[1,2]

1. University of Zurich
2. Swiss Institute of Bioinformatics

Keywords: Bioinformatics, visualisation

Webpages: https://bioconductor.org/packages/release/bioc/html/CrispRVariants.html, http://imlspenticton.uzh.ch:3838/CrispRVariantsLite/

CRISPR gene editing technologies have revolutionised molecular biology in the brief period since their development. A CRISPR gene editing experiment involves introducing a short DNA template into a cell together with a bacterially-derived enzyme capable of cutting DNA. The template-enzyme complex binds to the matching DNA of the host cell and cuts the host DNA. Mutations can be introduced during the repair of the cut, or a desired sequence can be added into the genome. In this way, genes and genetic control sequences can be precisely edited. Novel applications of CRISPR editing technologies are rapidly being developed. One way researchers assess editing efficiency is by sequencing the targeted DNA region. We developed an *R/Bioconductor* package **CrispRVariants** that allows researchers to evaluate and visualise sequences resulting from a gene editing experiment. We aimed to cater for a broad range of *R* programming abilities by providing a basic workflow whilst also allowing users the flexibility needed for novel experimental designs. The accompanying Shiny app **CrispRVariantsLite** makes analyses accessible to researchers with no *R* programming experience.

Poisson hierarchical model to analyze allele specific RNA-seq data

Ignacio Alvarez[1], Jarad Niemi[1] and Dan Nettleton[1]

1. Department of Statistics, Iowa State University

Keywords: Fully Bayesian, Hierarchical model, GPU, RNAseq

Fully Bayesian inference of high dimensional hierarchical models is computationally demanding. Usually, approximations like empirical Bayes or nested Laplace are used to obtain inference results. Parallel computing is a way to tackle down the computational intractability of this models, Landau and Niemi (2016) propose to use graphics processing units (GPU) to take advantage of the embarrassingly parallel nature of the MCMC algorithms in conditional independent hierarchical models, implemented in **fbseq** package (Landau 2016). We propose a hierarchical overdispersed Poisson model to study allelic gene expression in plants, fully Bayesian inference is obtained using **fbseq** package.

Allele specific expression (ASE) is a measure relative of the gene expression level of each gene copy in a multiploid genome. Biologically, ASE is relevant to explain process like hybrid vigor in plants or to design possible treatments for diseases in humans. In plant breeding, hybrids are developed to take advantage of the genetic phenomenon known as heterosis or hybrid vigor. Heterosis occurs when hybrid offspring possess superior levels of one or more traits relative to their inbred parents. Recent genomic studies suggest phenotypic heterosis may be explained by heterosis in the expression levels of key genes. Furthermore, one possible reason for the occurrence of heterosis are genes where two distinct alleles at a heterozygous locus are differentially expressed

ASE has some characteristics different from regular RNAseq expression: it is not available for every gene, it present bias towards one of the alleles (reference allele), and it has always a split-plot design. We present statistical methods for modelling ASE and detecting genes where differential allele expression. We propose a hierarchical overdispersed Poisson model that accommodates gene specific overdispersion, it has an internal measure of the reference allele bias, and use random effects to model the gene specific regression parameters. Simulation and real data analysis suggest the proposed model is a practical and powerful tool for the study of differential allele usage.

References

Landau, Will. 2016. *Fbseq: Fbseq*. https://github.com/wlandau/fbseq.

Landau, Will, and Jarad Niemi. 2016. “A fully Bayesian strategy for high-dimensional hierarchical modeling using massively parallel computing,” 1–27. http://arxiv.org/abs/1606.06659.

Automatic generation of item analysis report in ShinyItemAnalysis using Rmarkdown

Jakub Houdek[1], Patricia Martinkova[2]

1. Faculty of Informatics and Statistics, University of Economics, Prague
2. Institute of Computer Science, Czech Academy of Sciences

Keywords: Reports, Rmarkdown, shiny, ShinyItemAnalysis, psychometric analysis, item analysis

Webpages: https://shiny.cs.cas.cz/ShinyItemAnalysis/, https://cran.r-project.org/web/packages/ShinyItemAnalysis/index.html

Psychometric analysis is essential in development of any educational test. Package **ShinyItemAnalysis** (Martinkova, Drabinova, Leder, & Houdek, 2017) provides web interface for analysis of educational tests and their items. It provides a wide range of methods and ouputs, one of which is an automatically generated report.

This presentation focuses on report generation in **ShinyItemAnalysis** which allows users to easily and automatically generate reports for given dataset (package or user provided). Reports are created using **Rmarkdown** (Allaire et al., 2016) within **Shiny** (Chang, Cheng, Allaire, Xie, & McPherson, 2017) environment.

We also present challenges we met and future plans we have with report generation. We believe automatic report generation which is accessible for free and online may increase usage of psychometric analyses in test development.

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., … Hyndman, R. (2016). *Rmarkdown: Dynamic documents for r.* Retrieved from https://CRAN.R-project.org/package=rmarkdown

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). *Shiny: Web application framework for r.* Retrieved from https://CRAN.R-project.org/package=shiny

Martinkova, P., Drabinova, A., Leder, O., & Houdek, J. (2017). *ShinyItemAnalysis: Test and item analysis via shiny.* Retrieved from https://CRAN.R-project.org/package=ShinyItemAnalysis

Do Presidents Have Sentiments?

Salvino A. Salvaggio, PhD[1]

1. Qatar Science & Technology Park

Keywords: political science, sentiment analysis, visualization

Webpages:

Although data analysis has been fully incorporated in the literary studies field for over a decade, an in-depth investigation of the textual or verbal production of public administration with equal quantitative rigor has not occurred. This work relies on data analysis tools (in *R*) to explore all the Italian Presidents' New Year speeches from 1949 to 2015. The 67 speeches are analyzed through quantitative methods such as descriptive statistics (*exploratory data analysis*), natural language processing (*NLP*), sentiment analysis and opinion mining using *R* and *R Packages.* The aim is to use data science methodologies to enrich and enhance political studies.

- Descriptive statistics were used to quantify the way(s) each President speaks to the Nation. Amongst others, it allowed differentiating elocutionary styles: crisper (202 words/speech) or verbose (3,513 words/speech), direct (17 words/sentence) or convoluted (49 words/sentence), slow (95 words/minute) or fast (142 words/minute), as well as variations to means. When applied to the time series, the descriptive analysis shows the mutations of the elocutionary styles over time and the fact that they are not always in line with the zeitgeist.
- Natural language processing methods highlighted the frequency and associations of single or groups of words. This was useful to extract the features of the New Year speeches overall but also the main interests of each President (with the oldest President in the history of the Italian Republic being the most worried about the future of the young generation). Quantified examples are given for 7 themes: unemployment, work/job, youth, culture, terrorism, reform, and homeland. Absolute and relative frequencies of these themes were computed and compared to the average frequency of the same themes in the language overall for the same period. Supported by meaningful independence t-tests and confidence intervals, this approach showed the comparative evolution of the recurrence of the 7 topics. But it also showed it can be generalized to any theme.
- After having built a "sentiment dictionary", quantitative sentiment analysis and opinion mining have been applied to quantify the expression of ideas, opinions, and statements as positive or negative based on the wording. Relevant differences between Presidents emerge with, at the 2 extremes, President Pertini (18% positive sentiments against 9% negative) and President Gronchi (27% positive sentiments against 4.5% negative). Also, historical trends become more visible: towards more pessimism in the 1980s followed by a slightly stronger optimism in the 1990s and again more negative sentiments from 2000 onward. Sentiment analysis also made obvious that some Presidents built up their narratives following recurrent "sentiment/opinion patterns". The most evident case is President Napolitano that alternates good and bad news in such a specific manner that it becomes a pattern signature structuring most of his speeches.

The *data-oriented* approach adopted in this analysis has produced unique insights into the institution's textual production and its variation over time that add an extra dimension to the field of political studies.

R Packages used for this work: **changepoint**, **dplyr**, **ggplot2**, **koRpus**, **lubridate**, **RWeka**, **string**, **tm**, **wordcloud**.

Bayesian item response modeling: an application to universities admission tests

Javier Martínez[a], Irene García Mosquera[b]

[a]Department of Scientific Computing and Statistics, Universidad Simón Bolívar, Venezuela,
[b]Department of Mathematics and Informatics, Universitat de les Illes Balears, Spain

Abstract: We introduce a novel application that we have developed with R on the problems of estimating difficulties, discrimination and chance to guess right answers of the issues from the admission test to universities. Our application implements a Bayesian item response model, and we evaluated the performance of the model on a real data set. It consists on the responses that gave the students attending to the admission test to Universidad Simón Bolívar at Venezuela on 2012. Results were compared with those obtained through a standard method based on a classical statistical approach that we have also implemented on R. According to our results, Bayesian approach overcomes the classical one. In addition, we have also included in the application a generalized regression model Gamma with log link. That was used to forecast the performance of students as a function of the right responses to the issues having the highest score as assigned by our model. We have made a prospective analysis on the efficiency of that group of students and we will show that our predictions fit with the outputs.

Keywords: item response theory, Bayesian estimation, generalized linear model Gamma with log link, admission tests to Universities

Recency, Frequency, and Monetary Value in the 21st Century

Jim Porzak, DS4CI.org

Keywords: Recency, Frequency, Monetary, RFM, Marketing

Webpages: https://github.com/ds4ci/rfmr

Recency, Frequency, and Monetary Value (or simply RFM) is arguably the first use for "data-driven" marketing by direct mail practitioners in the second half of the last century. While the initial methods were crude, the basic concepts are still valid and in use today. In fact, Recency and/or Frequency often are important predictors in propensity to purchase, or churn, models being built today. Furthermore, these simple ideas are easily grasped my marketing practitioners making them popular in the marketing community.

The **rfmr** package provides functions for: 1) calculating RFM metrics at the individual customer level; 2) building customer segments from their RFM scores; 3) tracking segment transitions over time to trigger specific messaging and offers; and 4) visualizing these RFM metrics and clusters. Depending on its size, the customer database can be local within R or on an external relational data base.

Constructing Tests for Ads Quality Raters

Miles JNV[1], Lipman, PJ[1] and Schmitt, K[1]

1. Google, Inc, Los Angeles, California, 90008, USA

Keywords: psychometrics, test construction, ability testing

When users search the internet using Google, they are shown ads that are associated with their search keywords. These ads should be relevant to the search and useful to the user, and to this end Google uses human raters to evaluate the quality of ads shown on the search page (creative quality) and the page that the user is sent to (landing page quality). Test items consist of query-ad pairs, and are rated on a scale from -100 to +100, where ratings < -50 indicate that dissatisfaction is likely, -50 to 0 indicate dissatisfaction possible (below 0 considered bad), 0 through 50 satisfaction possible (considered neutral), and 51+ satisfaction likely (considered good).

To ensure that raters understand how to determine ad quality they are tested on a monthly basis. Raters are asked to rate items where the quality has been pre-determined, agreed upon by expert raters. Test items are interspersed with other evaluation items, and raters are not aware which items are test items.

This paper describes our analysis of exam data collected over a six month period, using structural equation modeling and item response theory.

We analyzed test results from the six months April to October, 2016. Factor analysis of a polychoric correlation matrix of items revealed that the test was not unidimensional, and suggested that evaluating good ads, neutral ads and bad ads were distinct skills. Constructing mean scores of these three skills, a longitudinal second order multiple-trait multiple method confirmatory factor analysis (using the **lavaan** package) was found to fit the data, and showed correlations between the three factors (good, bad, neutral) which ranged from -0.18 t0 0.30.

We examined item performance within factors using item response theory (using the **ltm** package), constructing item characteristic curves and total information curves. The analysis revealed that several items were problematic and should be removed, and that the range over which the test provided adequate reliability was somewhat restricted.

We then developed and evaluated a longer test with better psychometric properties. We describe how item response theory is used to scale newly developed tests to ensure that our quality standards are consistent over time.

CHMM: an R package for coupled Hidden Markov Models

Xiaoqiang Wang[1,2], Emilie Lebarbier[2], Stéphane Robin[2], and Julie Aubert[2]

Keywords: Coupled Hidden Markov Models, Variational approximation, Copy Number Variation.

Hidden Markov models (HMM) provide a natural statistical framework for the detection of the copy number variations (CNV) in genomics. In this context, a hidden markov process is associated to each individual in order to detect and to classify genomics regions in different states (typically, deletion, normal or amplification). Structural variations from different individuals may be dependent. It is the case in agronomy where varietal selection programm exists and species share a common phylogenetic past. We propose to take into account for these dependencies in the HMM model. When dealing with a large number of individuals, the exact inference of our coupled Hidden Markov model becomes intractable. We thus propose an approximate inference algorithm based on a variational approach. The **CHMM** *R* package performs the inference of these coupled Hidden-Markov Models. We show through simulations the performance of the proposed method and apply it to a datasets with 336 maize lines.

Predictive Modeling of Emergency Care in Older Adults: Comparison of Supervised Learning Algorithms in R

Op den Buijs J[1], Nikolova-Simons M[1], Fischer N[2], Golas S[2], Felsted J[2], Schertzer L[3], Agboola S[2]

1. *Philips Research, Eindhoven, the Netherlands*
2. *Partners Connected Health, Partners Healthcare, Boston, Massachusetts, United States*
3. *Philips Lifeline, Framingham, Massachusetts, United States*

Keywords: Accountable Care, Philips CareSage, Emergency Department, Medical Alert Service, Predictive Modeling

Webpages: https://www.lifeline.philips.com/business/caresage.html, http://connectedhealth.partners.org/, https://www.youtube.com/watch?v=9ytlU-7vzK0

Predictive analytics can be used to identify elderly patients at high risk for emergency care. In addition to use of Electronic Health Record (EHR) data collected before or at hospital discharge, non-hospital data may be useful for prediction of changes in risk outside of hospital settings. Medical Alert Services, i.e., services that enable older adults to get help in case of an emergency, collect such data while the elderly are living independently at home.

The objective of this study was to develop and validate predictive models of emergency care based on a combination of Medical Alert Service and EHR data in a population of elderly patients. In addition, a comparison of various supervised learning algorithms was performed to determine best performing methods across a variety of metrics.

Medical Alert Service and EHR data of 1,937 patients from a large healthcare organization were linked. Predictive models of emergency care in one year were developed using independent development and validation cohorts. Predictive model features were derived from patients' demographic information, two-year Medical Alert Service and clinical data. Supervised learning techniques used included logistic regression, random forest and boosted regression trees. Performance was evaluated by area under the receiver operator characteristic curve (AUC) and positive predictive value using the 90th percentile as a threshold (PPV@10%).

Using the full feature set ($n = 74$) for model development, AUC values and PPV@10% were highest for boosted regression in the validation cohort (AUC = 0.73, PPV@10% = 79%), and random forest reached similar performance (AUC = 0.72, PPV@10% = 78%). After a round of feature selection, logistic regression performance was at the same level of performance (AUC = 0.72, PPV@10% = 78%) with only 11 features. The most important features included the number of previous emergency incidents and falls at home, as well as previous outpatient encounters for urinary tract infection and atrial fibrillation.

Healthcare providers could benefit from our validated predictive models by estimating the risk of emergency care for individual patients and target timely preventive interventions to high risk patients. This could lead to overall improved patient experience, higher quality of care and more efficient resource utilization.

Importance sampling type correction of approximate MCMC for faster state space modelling

Jouni Helske

Department of Mathematics and Statistics, University of Jyväskylä

Keywords: Bayesian inference, state space models, time series, Markov chain Monte Carlo, Stan

Webpages: https://github.com/helske/stannis

State space models offer a flexible framework for time series analysis. Popular special cases include Gaussian structural time series models (Harvey 1989) and their non-Gaussian generalizations. This poster introduces a new *R* (R Core Team 2017) package **stannis** for Bayesian analysis of exponential family state space models where the state dynamics are Gaussian but the observational density is either Gaussian, Poisson, binomial, or negative binomial. The novel approach used by the **stannis** package is based on the combination of fast Gaussian approximation (Durbin and Koopman 2000), efficient No-U-Turn sampler (Hoffman and Gelman 2014) provided by *Stan* (Stan Development Team 2016), and the parallelisable importance sampling type corrected Markov chain Monte Carlo approach introduced by Vihola, Helske, and Franks (2016). We introduce the main modellings steps behind the **stannis** package, and illustrate the potential computational gains with simulated Poisson time series data.

References

Durbin, J., and S. J. Koopman. 2000. "Time Series Analysis of Non-Gaussian Observations Based on State Space Models from Both Classical and Bayesian Perspectives." *Journal of Royal Statistical Society B* 62: 3–56.

Harvey, A. C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press.

Hoffman, M. D., and A. Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15: 1593–1623.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2016. "RStan: The R Interface to Stan." http://mc-stan.org/.

Vihola, M., J. Helske, and J. Franks. 2016. "Importance sampling type correction of Markov chain Monte Carlo and exact approximations." Preprint 1609.02541v2. https://arxiv.org/abs/1609.02541.

R2GPU: A Simple *R* Interface for General Purpose Computing on NVIDIA Graphical Processing Units

Julio Olaya, Xueyuan Cao, and Stan Pounds

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA

Keywords: graphical processing unit, GPU, parallel computing

Webpage: https://www.stjude.org/pounds

Graphical processing units (GPUs) are specialized computing devices which were originally developed to accelearate graphics rendering applications that require intensive parallel processing of data. NVIDIA GPUs are widely available as plug-in expansion cards for servers or personal computers. GPUs consist of a large number of cores that rapidly process data in parallel in a manner analogous to the CPUs in a high-performance computing infrastructure. We have developed **R2GPU** software as a suite of *R* packages (**R2GPUbase**, **R2GPUmath**, **R2GPUprob**) that provide a simple *R* interface to general purpose computing on NVIDIA graphical processing units (GPUs). The **R2GPUbase** package defines the infrastructure to initiate interactions between an *R* session and a GPU device, transmit data between an *R* session and a GPU device, and transmit instructions from *R* to a GPU device. The **R2GPUmath** and **R2GPUdist** packages define mathematical operations and probability distributions. The **R2GPU** infrastructure is defined in a manner that provides two distinct practical advantages to the *R* programmer: (1) most of the *R* commands to perform operations on the GPU are syntactically identical to standard *R*, thereby minimizing the burden of learning a new set of *R* commands; and (2) the results of a GPU calculation may remain on the GPU for further processing by the GPU without intermediate GPU-CPU transmissions, thereby greatly reducing the overall time required to complete a series of operations on the GPU. In this way, the **R2GPU** packages provide *R* programmers the ability to easily modify existing packages to perform computationally intensive tasks on the GPU. Individual operations such as sorting large vectors or matrix multiplication have 50-1000 fold speed-up. As a real application, we used the **R2GPU** packages to develop a GPU implementation of a permutation-based procedure that evaluates the association of each of many genomic features with a phenotype. The GPU implementation obtained a 62-fold speed-up. The **R2GPU** packages provide *R* programmers an intuitive interface to develop GPU implementations of computationally intensive algorithms.

Predicting Long-Term Functional Outcome after Ankle Fracture

Juriaan van den Berg[1], *Bart Smeets*[2] *and Harm Hoekstra*[1,3]

1. *Katholieke Universiteit Leuven*
2. *dataroots*
3. *University Hospitals Leuven*

Keywords: AO44, Predictive Modelling, Trauma Care, Ankle Fracture

When it comes to data analysis, healthcare research has had a strong focus on being able to correctly explain the data at hand. However, over the last years, and in part due to evolutions in the field of machine learning, making predictions based on data has become a more frequent topic of discussion. Although explaining data and making predictions based on these data is strongly intertwined, they do differ in approach. While the former focusses on correctly understanding variable significance, the latter tries to optimise predictive performance.

In recent years, we have tried to gain a better understanding of patients' outcome after trauma care. Our most recent endeavours focus on exploring the use of predictive models to predict long-term functional outcome after AO type 44 ankle fractures (Müller 2014). Different machine learning models were trained and systematically benchmarked by leveraging the **mlr** package (Bischl et al. 2016).

We will present the achieved predictive performance, our considerations on practical applications and explain some of the difficulties we encountered along the way. On a more general note, we will discuss the value and risk of predicting long-term outcome, the current costly method of acquiring outcome related data and the relatively small sample sizes that this results in.

References

Bischl, Bernd, Michel Lang, Jakob Richter, Jakob Bossek, Leonard Judt, Tobias Kuehn, Erich Studerus, Lars Kotthoff, and Schiffner Julia. 2016. *Mlr: Machine Learning in R.* https://CRAN.R-project.org/package=mlr.

Müller, Maurice E. 2014. "Müller AO Classification of fractures-long bones."

Applications of Big Data methods in Finance: Index Tracking

Kamil Simka[1], *Luca Margaritella*[2]

1. Faculty of Economic Sciences, University of Warsaw
2. School of Business and Economics, Maastricht University

Keywords: Index tracking, Passive investment, L1-norm, Least Angle Regression, Quadratic optimization, Warsaw Stock Exchange, WIG, WIG20, mWIG40, sWIG80

Over the last decades big data revolution has influenced new research tech- niques in numerous fields of science. Lately it has been widely popularized also in economics. New methods have been introduced to solve complex prob- lems where the biggest concern is high-dimensionality. These issues mostly occur in micro- and macro-economics environments, where researchers ob- serve large number of variables of few observations. For instance, in macroe- conomics, when identifying the monetary transmission mechanism, one can access various indicators of the current economic state. However, such vari- ables are often quoted infrequently, say on quarterly basis, which can lead to problems when standard econometrics techniques are applied. See Bernanke (2004).

Such issues, however, do not occur normally in finance. Although, the num- ber of financial instruments quoted on stock exchanges is extremely large, high frequency data is easily available. Hence, the ”trap” represented by the curse of dimensionality does not relate to most financial settings. Nev- ertheless, some researchers have adopted high-dimensional methods for few finance-related problems. Index tracking is an example of such a setting. An equity index can be treated as a portfolio of stocks, weighted in a predefined manner. Although index construction is widely known to investors, it is of- ten more efficient to, instead of replicating an entire index, find an optimal sample that simply mimics its behavior.

In R we build an algorithm that, taking into account optimization con- straints, finds the most effective set of stocks that are used to track an index. The algorithm is based on the Least Angle Regression procedure of Efron, Hastie, Johnsone and Tibshirani (2004), however significant improvementsare proposed. Equity is a peculiar object, that requires specific assump- tions when being modeled. Trading involves costs that depend on market liquidity and brokerage commissions. Short sale (selling borrowed assets) on many stock exchanges is limited. Purchase fractional quantities of stocks is not possible. Although these constraints often do not apply to large invest- ment management companies, they are of particular concern to individual investors.

Additionally, we argue that index tracking is more of an ongoing process, rather than a point estimation problem. Instead of solving an optimization problem once, an index tracking strategy requires finding optimal portfolio for any point in time. After construction of a tracking portfolio, the proposed algorithm systematically examines if a replacement of portfolio’s constituents is optimal. Hereby, we propose a consistent investment strategy that can be applied readily by means of using the R software.

The algorithm is applied to tracking indexes of Polish stock exchange (WIG20, mWIG40, and sWIG80 market indexes,) giving a valid example of a practi- cal financial environment. Although, index tracking can be applied to any equity indexes, stock exchanges of developing countries are the most likely environment to apply sampling methods. The most developed exchanges pro- vide natural index tracking tools in form of Exchange Traded Funds (ETFs), investment funds that thanks to economies of scale efficiently track equity indexes. Such instruments are usually lacking on less developed exchanges, leaving investors with no other choice than independently construct tracking portfolios.

Teaching the Tidyverse in the Second Semester, Undergraduate Statistics Course

Kelly McConville

Keywords: statistics education, tidyverse, markdown, glmnet

The second semester, undergraduate statistics course has historically focused on modeling, with an emphasis on generalized linear models. To modernize my course, I wanted to satisfy two popular but conflicting ideas:

1. Teach the entire data analysis workflow (Wickham and Grolemund 2016), of which modeling is only one step.
2. Teach a more diverse set of models, especially statistical learning techniques.

But, how would I have more time to teach other aspects of analysis if I also wanted to cover new modeling techniques? In this talk, I present how I used the Tidyverse (Wickham 2016) to teach many steps of the data analysis workflow. I streamlined the process with packages such as **dplyr** and **magrittr** for data wrangling, **ggplot2** and **shiny** for data visualization, **broom** for tidy output, and **rmarkdown** for all assignments. This freed up class time to cover new modeling techniques, such as elastic-net penalized regression with the **glmnet** package.

References

Wickham, H. 2016. "Tidyverse." http://tidyverse.org/.

Wickham, H., and G. Grolemund. 2016. "R for Data Science." http://r4ds.had.co.nz/; O'Reilly Media.

Enlighten the past: five years of luminescence data analysis using R

Kreutzer S[1] and Burow C[2] and Dietze M[3] and Fuchs M C[4] and Fischer M[5] and Schmidt C[5]

Keywords: Luminescence, Chronology, Geoscience

Webpages: https://CRAN.R-project.org/package=Luminescence, http://www.r-luminescence.org

Enlighten the past: five years of luminescence data analysis using R

Geochronological research aims at deciphering and constraining palaeoenvironmental landscapes and processes. Luminescence dating is the method of choice for establishing chronologies by determining the last event of heating or sunlight exposure of natural mineral grains (e.g., quartz, feldspar).

Since 2012 we are developing data analysis tools based on *R* to support and enhance geochronological research on various levels with a particular focus on luminescence dating. Our contribution gives an overview of existing *R* packages to analyse luminescence data.

Additionally, we present the S4-object class structure implemented in the *R* package **Luminescence**, which is specifically tailored to deal with luminescence data. Our so-called RLum-object system enables a seamless data exchange across linked packages. The objects are carrying raw measurement data, as well as object specific metadata (e.g., the name of the creator function). Furthermore, by using unique identifiers, set at the time the object is created, changes in objects and applied methods can be tracked later on. Our object design paves the way for a very flexible and powerful data analysis, without diving deep into the *R* language itself.

Analyzing digital PCR data in R

Machteld Varewyck[1]

1. *Open Analytics NV*

Keywords: automated analysis, digital PCR, DNA amplification, R-package

Digital Polymerase Chain Reaction (dPCR) analysis has become a valuable and widely used alternative to real-time quantitative PCR analysis. Instead of analyzing a single sample, the sample is split into a large number of subreactions (e.g. 20 000), where each subreaction is amplified using typical quantitative PCR protocols.

Analysis of single channel digital PCR data basically consists of two steps: (1) Determining the proportion of negative (and positive) subreactions and (2) Given these proportions, estimating the concentration of target DNA. Assuming a Poisson distribution for the proportions, the second step entails a simple calculation. However, for the implementation of the first step there is no such clear consensus. We will therefore present and compare different methods implemented in *R*: Dreo et al. (2014), Jones et al. (2014), Trypsteen et al. (2015), Lievens et al. (2016).

These methods differ on how they define a threshold to distinguish between positive and negative subreactions, given the observed amplitudes (fluorescent intensity). Comparison of the statistical methods becomes especially interesting when some subreactions amplified only partly, i.e. in the presence of so-called 'rain'. With a single upload of the dPCR data, the *R*-package we present produces uniform output for each of the methods enabling direct comparison of the obtained concentrations.

References

Dreo, Tanja, Manca Pirc, Živa Ramšak, Jernej Pavšič, Mojca Milavec, Jana Žel, and Kristina Gruden. 2014. "Optimising Droplet Digital Pcr Analysis Approaches for Detection and Quantification of Bacteria: A Case Study of Fire Blight and Potato Brown Rot." *Analytical and Bioanalytical Chemistry* 406 (26). Springer: 6513–28.

Jones, Mathew, James Williams, Kathleen Gärtner, Rodney Phillips, Jacob Hurst, and John Frater. 2014. "Low Copy Target Detection by Droplet Digital Pcr Through Application of a Novel Open Access Bioinformatic Pipeline,'definetherain'." *Journal of Virological Methods* 202. Elsevier: 46–53.

Lievens, A, S Jacchia, D Kagkli, C Savini, and M Querci. 2016. "Measuring Digital Pcr Quality: Performance Parameters and Their Optimization." *PloS One* 11 (5). Public Library of Science: e0153317.

Trypsteen, Wim, Matthijs Vynck, Jan De Neve, Pawel Bonczkowski, Maja Kiselinova, Eva Malatinkova, Karen Vervisch, Olivier Thas, Linos Vandekerckhove, and Ward De Spiegelaere. 2015. "DdpcRquant: Threshold Determination for Single Channel Droplet Digital Pcr Experiments." *Analytical and Bioanalytical Chemistry* 407 (19). Springer: 5827–34.

grintar: A demonstration of reproducible analysis, visualization and distribution of ergometer exercise data

Marc Teunis[1], Jan-Willem Lankhaar[2,3], Shirley Kartaram[1],
Eric Schoen[4], Raymond Pieters[1,5]
1. Innovative Testing, HU University of Applied Sciences Utrecht
2. Digital Smart Services Research Group, HU University of Applied Sciences Utrecht
3. Institute for ICT, HU University of Applied Sciences Utrecht
4. TNO, Zeist
5. IRAS, Utrecht University

Keywords: reproducibility, data provenance, version control, exercise physiology, biomarker discovery

Webpages: https://github.com/uashogeschoolutrecht, http://www.innovativetesting.nl

Reproducibility of research results is essential for the progress of science. However, in many cases it has not kept pace with the recent explosion of analytical technology. While most modern analytical tools do provide means to stimulate reproducible analyses (e.g. versioning), these are often put into practice ineffectively. As a result, outcomes are often poorly reproducible. The **grintar** *R* package demonstrates how complying with a number of principles during analysis can greatly improve reproducibility.

The **grintar** package contains the raw and processed data and analysis and visualization methods from GRINTA!, a recent ergometer exercise study in 15 healthy volunteers. They underwent different cycle ergometer exercise regimens in order to identify biomarkers for recovery after exercise. Serum, urine and saliva samples were collected at baseline, during and after cycling and about 100 biological parameters were determined from each sample (publication in preparation). Sample analysis was conducted in six different laboratories.

For preprocessing and analysis, Ridge's *guerilla analytics* approach (Ridge 2014) was followed. This approach is based on seven principles: clarity, agility, simplicity, automation, data provenance, version control, knowledge consolidation and integrity of runs. Especially, data provenance (i.e. maintaining a link between the original data and the analysis outcome) is an important aspect of the approach.

In practice, guerilla analytics means: maintain a simple project folder structure, keep raw data as is (including filenames), use a data log, assign a unique ID to each data set, automate all preprocessing and analysis steps, use version control, use a work products log, assign a unique ID to each work product, use data builds and follow naming and coding style conventions. For coding style, Wickham's style (Wickham 2014) was followed.

All data preprocessing, analysis and visualization steps were automated and stored in the **grintar** package, as well as the raw and resulting data sets. In addition, the package was fully documented. After scientific publication, it will be published on GitHub to be used for educational and scientific purposes.

The **grintar** package demonstrates that the guerilla analytics approach provides practical and useful guidelines for reproducible data analysis.

References

Ridge, Enda. 2014. *Guerrilla Analytics: A Practical Approach to Working with Data.* Morgan Kaufmann.

Wickham, Hadley. 2014. *Advanced R.* CRC Press.

Censoring in random effects models

Matthias Kuhn[1] *and Ingo Roeder*[1,2]

1. Institute for Medical Informatics and Biometry (IMB), Faculty of Medicine Carl Gustav Carus
Technische Universität Dresden, Dresden, Germany

2. National Center for Tumor Diseases, Partner Site Dresden, Germany

Keywords: censoring, random effects model

Webpages: https://github.com/lenz99-/lme4cens

Random effects models are one option in regression analysis when the data has a hierarchical structure, i.e. when the observations are *not* independent and identically distributed. Random effects models account for the induced correlation by assigning zero-centered normally distributed random variables to different hierarchy levels in the data and by predicting their value during the fitting process. As an example, repeated measurements in a longitudinal study are nested within subjects and e.g. a simple random effects model would provide for a random intercept per subject. In the *R* world, random effects models are handled – among others – in the **lme4** package. Given a data sample, the parameters of a random effects model are estimated in **lme4** via (restricted) maximum likelihood.

If a response is not known exactly but only to have occurred within a certain interval we speak of censoring. It is a well-known concept from time-to-event analysis but censoring also occurs e.g. when the response is a concentration and the measuring device has a lower detection limit. There exist *R*-implementations for regression analysis with censored response, most notably the **survival**-package with its `survreg`-function for parametric regression models. Regarding random effects models, there are custom *R*-packages (e.g. **censReg**, Henningsen (2010)) that support censoring as well, at least for simple random intercept models.

The **lme4**-package is modular and developers are encouraged to reuse functionality for model enhancements or specializations (Bates et al. 2015). We have developed our *R*-package **lme4cens** that builds on **lme4** to fit simple random effects models with a censored response. In particular, the re-use of the formula-module facilitates model specification. The censoring information is encoded via **survival**'s `Surv`-object that allows for a flexible specification of (a combination of) left-, right- and interval-censored responses with varying censoring levels.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software*. https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

Henningsen, Arne. 2010. "censReg: Censored Regression (Tobit) Models. R Package." http://cran.r-project.org/package=censReg.

CONS an R Based Graphical Interface to Perform Consonance Analysis

N. Sofia Huerta-Pacheco[1] *Victor M. Aguirre*[2] *and M. Teresa Lopez* [3]

1. Universidad Veracruzana
2. Statistics Department, ITAM
3. Delphi Intelligence

Keywords: Sensory evaluation, Panel performance, PCA

Webpages: https://CRAN.R-project.org/package=CONS

Consonance Analysis is a useful numerical and graphical exploratory approach for evaluating the consistency of the measurements and the panel of people involved in sensory evaluation. It makes use of several uni and multivariate techniques, particularly Principal Components Analysis. This paper shows the implementation of this procedure in a graphical interface R package named CONS.

References

Dijksterhuis GB (1995). “Assessing Panel Consonance”. Food Quality and Preference, **6**(1):714. URL http://dx.doi.org/10.1016/0950-3293(94)P4207-M

Dijksterhuis GB (2004). “Multivariate data Analysis in Sensory and Consumer Science.” Food and Nutrition Press, Inc. USA, Trumbull Connecticut.

Error model estimation for dynamic models by maximum-likelihood methods

Fehling-Kaschek M[1], Mader W[1], Rosenblatt M[1], Timmer J[1] and Kaschek D[1]

1. *Institute of physics, Freiburg University, Germany*

Keywords: dynamic modeling, error modeling, parameter estimation

Webpages: https://github.com/dkaschek/dMod

Mathematical modeling has become an established approach in systems biology to gain information about underlying processes. Our package **dMod** was developed for dynamic modeling based on ordinary differential equations. The model parameters are estimated via the maximum likelihood method, in general requiring time-resolved data. Depending on the measurement technique, taking data points is time-consuming and expensive. Therefore, the modeler is often confronted with the problem of low number of replicates from which uncertainties need to be estimated reliably.

Error models provide a way to pool replicate measurements from different time-points and conditions to estimate the contributions from different error sources. Here, two complementary maximum-likelihood approaches to identify error model parameters are implemented: (1) in a standalone fit from mean-variance tuples and (2) from model residuals in a combined estimation with the dynamic model. Advantages and disadvantages of both approaches are discussed and usecases presented.

AmyloGram: prediction of amyloid sequences in R

Michał Burdukiewicz[1], Piotr Sobczyk[2], Stefan Rödiger[3], Anna Duda-Madej[4], Marlena Gąsior-Głogowska[1], Paweł Mackiewicz[1] and Małgorzata Kotulska[1]

1. University of Wrocław
2. Wrocław University of Science and Technology
3. Brandenburg University of Technology Cottbus-Senftenberg
4. Wrocław Medical University

Keywords: Amyloids, n-grams, machine learning, random forest

Webpages: CRAN | GitHub | Shiny

Amyloids are proteins associated with important clinical disorders (e.g., Alzheimer's or Creutzfeldt-Jakob's diseases). Despite their great diversity, all amyloid proteins can undergo their aggregation initiated by 6- to 15-residue segments. The aggregation propensity seems not to depend on the exact amino acid residues, but rather on their physicochemical properties. Therefore, we created a model of amyloidogenicity incorporating this knowledge.

We created 524,284 reduced amino acid alphabets based on diversified combinations of the physicochemical properties of amino acid residues. Using very fast implementation of the random forest classifier from the **ranger** package we cross-validated all reduced alphabets and identified one that provided the best discrimination between amyloids and non-amyloids. Our feature selection method found 65 motifs that are the most relevant to the discrimination of amyloid and non-amyloid sequences.

The reduction of amino acid alphabet turned out to be very efficient because most of the predictors based on them outperformed those trained on the full amino acid alphabet. This result confirmed our assumption on the role of more general amino acid properties in amyloidogenicity. Thanks to the relatively large number of evaluated alphabets, we were able to confirm the significance of amino acid flexibility in the amyloid aggregation. Among 65 most informative amino acid motifs identified during the analysis, 15 were independently confirmed in experimental studies [1]. The best-performing predictor, **AmyloGram** [2], was benchmarked against other tools for amyloid peptides detection using an external dataset. Our software obtained the highest values of performance measures (AUC: 0.90, Matthews correlation coefficient: 0.63).

AmyloGram was futher used to predict properties of amyloid sequences from the AmyLoad database [3]. We choose 24 hexapeptides which has the opposite annotation and the prediction of **AmyloGram** (12 hexapeptides annoted in the database as amyloidogenic and identified by **AmyloGram** as non-amyloidogenic and 12 hexapeptides annoted in the database as non-amyloidogenic and identified by **AmyloGram** as amyloidogenic). After the experimental validation with Fourier transform infrared spectroscopy we found that 13 peptides have apparently wrong annotation in database, which implies that model of **AmyloGram** is sensitive enough to identify false positives and false negatives in already known data.

Our analysis not only confirmed that amyloidogenicity depends on the general physicochemical properties of proteins, but also revealed which features are the most relevant to the initiation of amyloid aggregation. In addition, our framework identified amyloidogenicity-related amino acid motifs, which were previously confirmed experimentally. Aside from creation of the interpretative model of amyloidogenicity, we have also developed an accurate predictor of amyloids, **AmyloGram**.

References

1. López De La Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, Hoenger A, et al. De novo designed peptide-based amyloid fibrils. PNAS. 2002;99:16052–7.

2. Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. Prediction of amyloidogenicity based on the n-gram analysis. PeerJ Preprints; 2016. https://peerj.com/preprints/2390.

3. Wozniak PP, Kotulska M. AmyLoad: Website dedicated to amyloidogenic protein fragments. Bioinformatics (Oxford, England). 2015;31:3395–7.

Zero-overhead R and C/C++ integration with FastR

Mick Jordan and Lukas Stadler

Oracle Labs

Keywords: Java, C, C++, performance

Webpages: https://github.com/graalvm/fastr

FastR is an alternative implementation of the *R* language for statistical computing. It is based on the **Truffle** framework, which provides building blocks for language implementations and solves problems such as language interoperability at a fundamental level. Additionally, the **Graal** compiler provides Just-In-Time compilation from R to native code.

Traditionally, C and C++ are often used to improve performance for *R* applications and packages. While this is usually not necessary when using **FastR**, because it can run *R* code at near-native performance, there is a large corpus of existing code that implements critical pieces of functionality in native code. Alternative implementations of *R* need to simulate the R native API, which is a complex API that exposes many implementation details. They spend significant effort and performance overhead to simulate the API, and there is a compilation and optimization barrier between languages.

FastR can employ the Truffle framework to run native code, available as LLVM bitcode, inside the optimization scope of the polyglot environment, and thus have it integrated with no optimization and integration barriers. This talk briefly introduces the technical background and showcases how **FastR** can use fine-grained integration with native languages to gain unprecedented levels of performance for polyglot applications.

Genome-wide retroviral integration mapping in agressive T-cell leukemia

Rosewick Nicolas[1,2,†], *Hahaut Vincent* [1,†], *Artesi Maria* [1], *Durkin Keith* [1], *Marçais Ambroise* [3], *Griebel Philip* [4], *Arsic Natasa* [4], *Burny Arsène* [2], *Georges Michel* [1] *and Van den Broeke Anne* [1,2]

1. Université de Liège (ULg), Liège, Belgium

2. Université Libre de Bruxelles (ULB), Brussels, Belgium

3. Hôpital Universitaire Necker, Université René Descartes, Paris, France

4. University of Saskatchewan, Saskatoon, Canada

[†]Presenting authors

Contact: nrosewic@ulb.ac.be ; vincent.hahaut@ulg.ac.be

Keywords: Bioinformatics, Retrovirus, Genome, Integration, DNA sequencing

Among the genetic and environmental causes of cancer, viral infection is responsible for about 10 to 15% of human cancers worldwide [1]. Retroviruses are part of a specific category of viruses that have the capacity to integrate into the host genome. Large body of evidence shows the importance of the genomic localisation of the proviral integration site within the host genome and its potential role in cancer onset. Recent advances in ultra-deep DNA sequencing allow the genome-wide analysis of the integration landscape of several retrovirus classes, contributing to a better understanding of virus-host interactions and their role in pathogenesis.

In this work we focused on Human T cell leukemia Virus-1 (HTLV-1), a retrovirus that infects T lymphocytes and provokes an aggressive T-Cell Leukemia in ~5% of infected individuals after a long asymptomatic period (several decades). It is estimated that 10 to 20 million people worldwide are infected by HTLV-1. We used an experimental model of infection of sheep with the closely related Bovine Leukemia Virus (BLV). We analysed longitudinal samples from before infection to the acute aggressive leukemic stage, enabling to explore the dynamics and progression of infected clones during tumor progression.

Recently our group developed an improved high-throughput sequencing (HTS) method to map BLV/HTLV-1 integration sites and quantify the abundance of the corresponding clones [2]. Using a custom *R* pipeline we fine-mapped the positions of several hundred thousand integration sites in both HTLV-1 and BLV infected individuals. Simulation-based integration hotspot analysis revealed non-randomness of viral integration in tumors as well as in asymptomatic samples isolated from BLV-infected sheep. Their genomic distribution was significantly biased towards cancer driver genes arguing in favour of interactions between the provirus and host genes located in its vicinity. Altogether using a novel deep DNA sequencing approach combined to advanced *R*-based analysis, we were able to highlight an unexpected new role for HTLV-1 and BLV integration in the context of tumor onset and progression.

References

1. Moore PS, Chang Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. Nature Reviews Cancer. 2010;10:878–89. doi:10.1038/nrc2961.

2. Rosewick N, Durkin K, Artesi M, Marçais A, Hahaut V, Griebel P, et al. Cis-perturbation of cancer drivers by the HTLV-1/BLV proviruses is an early determinant of leukemogenesis. NatCommun. 2017.

startR - Retrieving multidimensional distributed data sets in R

Nicolau Manubens[1]
1. Barcelona Supercomputing Center (BSC)

Keywords: distributed data sets, multi-dimensional, subset, retrieve, homogenize, arrange

Data retrieval and alignment is the first step in data analysis, and is often highly complex and time-consuming. This is especially crucial in the era of Big Data, where large multidimensional data sets from diverse sources need to be combined and processed. In this context, the Divide and Conquer technique is indispensable. This talk introduces startR (Subset, Transform, Arrange and ReTrieve multi-dimensional subsets in R), an R project started at BSC with the aim to develop a tool that allows the user to automatically retrieve, homogenize and align subsets of multidimensional distributed data sets. The talk includes an explanation of the current features of the package and of its intended role in a Big Data workflow. startR is an open source project that is open to external collaboration and funding, and will continuously evolve to support as many data set formats as possible while maximizing its efficiency.

The causal impact of algorithmic trading on market quality

Nidhi Aggarwal[1] *and Susan Thomas*[1]

1. *Finance Research Group, Indira Gandhi Institute of Development Research*

Keywords: Algorithmic trading, matching, difference-in- difference, flash crashes, market quality

Webpages: http://ifrogs.org/releases/ThomasAggarwal2014_algorithmicTradingImpact.html

The causal impact of algorithmic trading (AT) on market quality is difficult to establish due to endogeneity bias. We address this problem by using the introduction of co-location as an exogenous event after which AT increased. Matching techniques are used to identify a matched set of firms with high and low AT to estimate causal impact. We find that securities with higher AT have lower liquidity costs, order imbalance, and price volatility. We find new evidence that high AT is not associated with higher intraday liquidity risk or higher incidence of extreme intraday price movements.

The E-learning System for Linear Models; The >eR-Biostat Initiative for Developing Countries

Nolen Joy Perualila[1], *Ziv Shkedy*[1,2], *Khangelani Zuma*[3], *Legesse Debusho*[4] *and Adetayo Kasim*[2,5]

1. Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, 3590 Diepenbeek, Belgium
2. Department of Epidemiology and Biostatistics, Gonder University, Ethiopia
3. Human Sciences Research Council (HSRC), PRETORIA, South Africa
4. The University of South Africa (UNISA), PRETORIA, South Africa
5. Wolfson Research Institute for Health and Wellbeing, Durham University, Durham

Keywords: Linear Models, The >eR-BioStat initiative, Developing countries, Master programs, The E-learning system using *R*.

Linear models are a central tool in data analysis in any scientific discipline and in particular in biostatistics applications. In the poster we present a new linear models course which was developed as a part of the **>eR-Biostat** Initiative. The linear models course provides online materials for master students in biostatistics/statistics in developing countries. The materials developed for the linear model course are *R* oriented and publicly available.

Several types of course materials are presented: (1) note for the course (the book "Practical Regression and Anova using R" by Julian J. Faraway which is available online in https://cran.r-project.org/doc/contrib/), (2) slides for the course, (3) *R* programs, ready to use, which contain all data and *R* codes for all the examples and illustrations discussed in the course and (4) homework assignments and exam.

In the poster we present the different types of course materials for a specific module in the linear models course (parameters estimation for linear regression) and present all the materials for the course which are currently available online in the **>eR-Biostat** website.

Estimating health service accessibility in Guatemala

O. de León
Centro de Estudios en Salud, Universidad del Valle de Guatemala

Keywords: Health, accessibility, distance, equity, GIS

Webpages: https://github.com/odeleongt/useR_2017/tree/master/01_gt_health_service_access

Guatemala (Central America) is a developing country with wide disparity issues regarding income and basic needs coverage (The World Bank 2017). Primary health service facilities are available in Guatemala over the whole country at a reasonable geographic distribution (Smiley 2012), in principle allowing for equitative access to health care for all the population. Physical accessibility to health services has been evaluated previously for some areas of the country (Annis 1981; Owen, Obregón, and Jacobsen 2010), but several shortcomings preclude the availability and usefulness of this information.

Physical accessibility to services is usually evaluated based on distance from the communities to the health services, either over a shortest path or through a known network of possible routes (Delamater 2013). Theses analysis topically require a considerable amount of analysis effort, and the use of traditional analysis tools (i.e. point-and-click style analysis software) has limited previous efforts to a few regions of the country (Annis 1981; Owen, Obregón, and Jacobsen 2010). Previous analysis have not taken into account altitude changes in the routing, which heavily influences transportation costs. Given the economic disparity in the country, altitude change should be considered to provide a better assessment of physical accessibility to health services.

This work relies on the use of R (R Core Team 2017) to provide an extensible analysis template which allows to update physical accessibility estimates, include new areas for which information is now available, and efficiently use currently available altitude data (Fujisada, Urai, and Iwasaki 2012). Usability is improved by taking advantage of recent developments in spatial data use (Pebesma 2017) and visualization (Cheng, Karambelkar, and Xie 2017) in R to improve ease of analysis and provide more relevant result dissemination.

References

Annis, Sheldon. 1981. "Physical Access and Utilization of Health Services in Rural Guatemala." *Social Science & Medicine. Part D: Medical Geography* 15 (4): 515–23. doi:http://dx.doi.org/10.1016/0160-8002(81)90046-0.

Cheng, Joe, Bhaskar Karambelkar, and Yihui Xie. 2017. *Leaflet: Create Interactive Web Maps with the Javascript 'Leaflet' Library*. https://CRAN.R-project.org/package=leaflet.

Delamater, Paul L. 2013. "Spatial Accessibility in Suboptimally Configured Health Care Systems: A Modified Two-Step Floating Catchment Area (M2sfca) Metric." *Health & Place* 24: 30–43. doi:http://dx.doi.org/10.1016/j.healthplace.2013.07.012.

Fujisada, Hiroyuki, Minoru Urai, and Akira Iwasaki. 2012. "Technical Methodology for Aster Global Dem." *IEEE Transactions on Geoscience and Remote Sensing* 50 (10). IEEE: 3725–36.

Owen, Karen K., Elizabeth J. Obregón, and Kathryn H. Jacobsen. 2010. "A Geographic Analysis of Access to Health Services in Rural Guatemala." *International Health* 2 (2): 143. doi:10.1016/j.inhe.2010.03.002.

Pebesma, Edzer. 2017. *Sf: Simple Features for R*. https://CRAN.R-project.org/package=sf.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Smiley, Anne. 2012. *Educational Quality Improvement Program: Nine Years of Experience in Education Policy, Systems, and Management*. USAID.

The World Bank. 2017. "Guatemala: Country Profile." The World Bank. September. http://www.worldbank.org/en/country/guatemala/overview.

BioMAn$^{\text{TM}}$: a user-friendly interface for targeted metagenomic data visualization and analysis

Pauline Vaissié[1], Christophe Camus[1], Yao Amouzou[1], Thomas Carton[1], Sophie Le Fresne-Languille[1], Françoise Le Vacon[1], Murielle Cazaubiel[1] and Sébastien Leuillet[1]

1. Biofortis Mérieux NutriSciences

Keywords: Metagenomics, Data visualization, Statistical analysis, Web platform

With the recent advances in the field of next-generation sequencing (NGS), metagenomics allow to explore the biodiversity of microbial ecosystems or microbiota. Dedicated bioinformatic pipeline focusing on targeted metagenomics (16S rRNA) provides to biologists the bacterial composition in OTUs (Operational Taxonomic Units) of the samples. Nevertheless, NGS produces massive data which requires substantial computer processing to extract information. Faced with this large amount of data, their visualization and appropriate statistical analysis are essential for scientists to adequately explore and interpret their experiments. In this context, we have developed an *R* [1] and *Shiny* [2] web based platform called BioMAn$^{\text{TM}}$ (Biofortis Metagenomics Analysis) which mixes the statistical power of dedicated R packages (**metagenomeSeq**, **mixOmics**, . . .) with a user-friendly web design. This interface allows users to interactively look into their project by manipulating, filtering or gathering information for further interpretation or communications purposes. Focusing on a subgroup of samples is made very easy by the integration of metadata table (information on samples such as experimental conditions). The core of the tool is focused on data visualization, which offers the possibility to depict taxonomic composition throughout several graphical interactive representations such as barplots, boxplots, heatmaps, Krona [3] or hierarchical trees. BioMAn$^{\text{TM}}$ also provides information (tables and graphs) about diversity indices to help users in the interpretation of results. This turnkey product is an easy way for scientists to conduct ordination analysis such as PCoA with a lot of customizable graphical and analytical options. The platform can also be used to run specific statistical analysis like discriminant analysis (LDA and FDA). Other statistical approaches are currently being added to the application (PERMANOVA/ANOSIM, differential analysis. . .) in order to create the fullest possible metagenomic toolbox. During the process, user can easily retrieve objects by downloading them in high quality or by inserting them one by one into a custom PowerPoint template. BioMAn$^{\text{TM}}$ is deployed on a Shiny Server Pro, implemented by a secure health data hosting provider according to the French regulatory requirements, to protect the confidentiality, integrity and availability of patient and user data.

References

[1] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[2] 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

[3] Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. "Interactive metagenomic visualization in a Web browser." BMC bioinformatics 12.1 (2011): 385.

useR! 2017 Poster Presentation

Visualizing Meeting Data in Shiny

Phoebe Wong

Legendary Applied Analytics

Abstract: Have you ever wondered how much of your work time was sitting in meetings and especially the ones that you do not need to be in? You are not alone. According to a Microsoft survey ("Survey Finds Workers Average Only Three Productive Days Per Week" 2005), U.S. workers reported spending 5.5 hours per week in meetings while 71 percent of them believe that meetings aren't productive. Another survey by Salary.com ("2014 Wasting Time at Work Survey" 2014) shows that 24% US workers surveyed think the biggest time-wasting activity at work was "having too many meetings/conference calls". Understanding how a worker's time is being used at work is important to the organization, managers and the worker him/herself.

Microsoft Outlook is commonly used in many workplaces for their email and calendar services. This Shiny application can analyze and visualize your meeting history in Outlook to provide a data-informed understanding of your meetings over time. Specifically, the application allows you to visualize your meeting history, with a user-specified range of dates, by providing an interactive heatmap of your meeting frequency, broken down by hours and days. It also analyzes and displays a network of the meeting attendees from your meeting history on an interactive network graph using D3.js and a static network graph using **igraph**. In addition, it also provides a table of your meetings using **data.table**, basic summary statistics with some frequency plots using **ggplot2**. The application allows a better understanding of how work time is spent and can help with making data-driven decisions for better resource allocation.

Keywords: Shiny, D3.js, igraph, plotly, Microsoft Outlook

References

"2014 Wasting Time at Work Survey." 2014. http://www.salary.com/2014-wasting-time-at-work/slide/6/.

"Survey Finds Workers Average Only Three Productive Days Per Week." 2005. https://news.microsoft.com/2005/03/15/survey-finds-workers-average-only-three-productive-days-per-week/#qHzMPOLe4TU4BuyR.97.

Modelling Spatial Point Data in R - sample applications

Piotr Cwiakowski[1] *and Piotr Wojcik*[1]

Faculty of Economic Scienses, University of Warsaw

Keywords: spatial econometrics, spatial point, local convergence, real estate valuation

Localization of observation is a crucial factor in many discipline, for example in epidemiology, biology, economics or regional sciences. Spatial point data are represented on maps as points so it should be precisely individually geolocalized. Modelling spatial points raises certain problems and methodology is in the development stage (Baddeley and Turner 2006, Baddeley (2009)). The first issue is abundance of data – currently available spatial methods are insufficient and slow – which does not fit available datasets and needs. Another problem is the definition of neighbourhood for point data. They are converted to polygons using so called Voronoi polygons (Okabe 1992), so that each point has a polygon and each polygon has one point. However, this neighbourhood structure is changing after modification/addition/removal of any observation in the dataset. Thus the spatial weights matrix, crucial in spatial modelling, is unstable in the training and testing set which makes prediction very difficult.

In our presentation we want to show sample applications of spatial point modelling, referring to valuation of real estates in Polish capital, Warsaw (static data) and investigating local convergence of educational achievements on the level of schools in Poland (dynamic data).

References

Baddeley, Adrian. 2009. "Analysing Spatial Point Patterns in 'R'."

Baddeley, Adrian, and Rolf Turner. 2006. "Modelling Spatial Point Patterns in R." In *Case Studies in Spatial Point Process Modeling*, 23–74. Springer.

Okabe, Atsuyuki. 1992. *Spatial Tessellations*. Wiley Online Library.

biogram: n-gram analysis of biological sequences in R

Piotr Sobczyk[1], *Chris Lauber*[2], *Pawel Mackiewicz*[3] *and Michal Burdukiewicz*[3]

1. Wroclaw University of Science and Technology, Faculty of Pure and Applied Mathematics
2. Technical University of Dresden, Institute for Medical Informatics and Biometry
3. University of Wroclaw, Department of Genomics

Keywords: Proteins, n-grams, machine learning, feature selection

Webpages: CRAN | Hsmm

n-grams (k-mers) are vectors of n characters derived from input sequences. Originally developed for natural language processing, they are also widely used in genomics, transcriptomics and proteomics. Despite continuous interest in sequence analysis, there are only a few tools tailored for comparative n-gram studies. Moreover, they often do not contain efficient feature-filtering methods, which severely hampers their use.

To facilitate comprehensive analysis of n-grams, we created **biogram**. Aside from essential functionalities, like efficient data storage, we also implemented a novel feature selection method. Commonly permutation tests are used for filtering important n-grams to find if an occurrence of n-gram and a value of a target are independent. However, exhaustive testing of permutations is computationally expensive and this often becomes a limiting factor. Therefore, we developed the Quick Permutation Test (QuiPT) which uses criteria such as information gained to choose the most discriminating features, without requiring exhaustive testing,

Several studies have highlighted that 3D protein structure depends not only on the exact sequence of amino acids but also on their general physicochemical properties. Therefore, a reduced amino acid alphabet which represents certain subgroups of amino acids, can still retain the information about the protein folding [1]. Since the function of proteins often depends on their structure, reduced alphabets can be used in prediction of protein roles. The reduction of the alphabet reduces the number of possible input variables (n-grams) streamlining development of the model and allow extraction of more straightforward decision rules.

The generation of reduced amino acid alphabet is facilited through implemented genetic algorithm [2]. It was modified to work faster and produce more adequate reduced alphabets. We also added distance measures for comparison of reducing alphabet, including similiarity index [3] and our own encoding distance.

We showcase the n-gram analysis in the problem of signal peptide prediction. Signal peptides are short subsequences present at the N-terminal of proteins. They serve a role similar to the postal code directing proteins to the endomembrane system and in consequence facilitating their export outside the cell. Here we implemented n-grams as hidden states in a hidden semi-Markov model (HSMM) of signal peptides called signalHsmm. This approach yields the largest AUC value (AUC = 0.98) in comparison to other software. Another advantage of the n-gram approach is universality of decision rules. Malaria-causing parasites from Plasmodium genus have slightly different amino acid composition of proteins than other organisms, which hinders prediction of their signal peptides in all models using the full amino acid alphabet. The usage of a reduced amino acid alphabet allows signalHsmm to employ universal decision rules, appropriate for proteins belonging to both malaria parasites and other organisms. Thanks to that signalHsmm recognizes signal peptides of malaria parasites with AUC = 0.92, more than other programs that cannot exceed AUC = 0.84.

References

1. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Engineering. 2000;13:149–52. doi:10.1093/protein/13.3.149.

2. Lenckowski J, Walczak K. Simplifying Amino Acid Alphabets Using a Genetic Algorithm and Sequence Alignment. In: Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics. Springer, Berlin, Heidelberg; 2007. pp. 122–31. https://link.springer.com/chapter/10.1007/978-3-540-71783-6_12.

3. Stephenson JD, Freeland SJ. Unearthing the root of amino acid similarity. Journal of Molecular Evolution. 2013;77:159–69.

Modelling distribution dynamics in R. Application to convergence analysis on a local level

Piotr W?jcik [1]

1. *University of Warsaw, Faculty of Economic Sciences*

Keywords: distribution dynamics, transition matrix, conditional kernel density estimation, ergodic distribution

When modelling different socio-economic phenomena (e.g. income, educational achievements, unemployment rate, political preferences, consumption of ice cream, etc.) one is often interested how much they are diversified in the sample and how their diversification changes over time – see e.g. Magrini (2009). The simplest approach concentrates on calculating a single measure of dispersion (e.g. coefficient of variation, Gini, Theil coefficients, etc.) and comparing it's values for subsequent periods.

But a single measure does not tell anything about the diversity within the distribution. So another popular approach takes into account the whole distribution by comparing histograms or unidimensional kernel density estimates in particular periods. However, this still tells nothing about the mobility within the distribution and does not allow for formulation of some kind of predictions (i.e. ergodic distributions).

This is possible when the distribution dynamics is modelled with the use of transition matrices (which requires discretization of the distribution) or conditional kernel density estimates as first proposed by Quah (1996).

This purpose of this presentation is to show how different approaches of modelling distribution dynamics can be applied in R, with a particular focus on transition matrices and conditional kernel density estimates. We will present our R based application of recently developed methodology allowing to summarize a two-dimensional conditional kernel density surface with the (univariate) ergodic distribution – see Gerolimetto and Magrini (2017).

We will also present readable and attractive ways of visualization of estimation results. Practical examples will be shown on simulated and real spatial data.

References

Gerolimetto, Margherita, and Stefano Magrini. 2017. "A Novel Look at Long-Run Convergence Dynamics in the United States." *International Regional Science Review* 40 (3).

Magrini, Stefano. 2009. "Why Should We Analyse Convergence Through the Distribution Dynamics Approach?" *Science Regionali* 8: 5–34.

Quah, Danny. 1996. "Twin Peaks: Growth and Convergence in Models Distribution Dynamics." *Economic Journal* 106: 1045–55.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability. Londyn: Chapman; Hall.

Zambom, A.Z., and R. Dias. 2012. "A Review of Kernel Density Estimation with Applications to Econometrics." Discussion Paper arXiv:1212.2812.

Multivariate ordinal regression models using the R package MultOrd

Rainer Hirk[1], *Kurt Hornik*[1] *and Laura Vana*[1]

1. WU Vienna University of Economics and Business

Keywords: Composite likelihood, Multivariate ordered logit, Multivariate ordered probit, R package

The R package MultOrd implements composite likelihood estimation in the class of multivariate ordinal regression models. A flexible modelling framework for cross-sectional as well as longitudinal observations is set up, which allows different error structures. Two different link functions are employed, by assuming a multivariate normal and a multivariate logistic distribution for the latent variables underlying the ordinal outcomes. In order to deal with the issue that absolute location and absolute scale are not identifiable in ordinal models, several restrictions on the parameter space are imposed either by fixing location parameters and/or by restricting the full covariance matrix to be a correlation matrix. In addition, constraints on both coefficient as well as threshold parameters can be imposed. Standard errors are calculated using the Godambe information matrix.

The R Journal: status

Roger Bivand[1]

1. Department of Economics, Norwegian School of Economics

Keywords: Review process, R documentation, article publication, author status

Webpages: https://journal.R-project.org

As the current responsible editor of **The R Journal**, I have the benefit of not only having to think about how the journal can best contribute to the R community, but also having access to the historical submission record. If one asks the right questions, it is possible to see some of what has been going on since the journal was created from the earlier **R News** in 2009.

This presentation will describe where we are now in the light of the last eight years, with numbers of submissions rising somewhat less fast than the CRAN package count, and what reverse dependencies may be present, for example with CRAN Task Views. Innovations in the interface to the journal website may increase discoverability, but there are clear opportunities for better propagation of journal content. Work is in train to use DOI linking, and other changes should be implemented before useR! and will be described.

Alternatives to the **akima** package

Roger Bivand[1] *and Albrecht Gebhardt*[2]

1. *Department of Economics, Norwegian School of Economics*
2. *Institut für Statistik, Alpen-Adria-Universität Klagenfurt*

Keywords: 2D interpolation, 2D extrapolation, 2D surface approximation, surface visualization

Webpages: https://CRAN.R-project.org/package=akima, https://CRAN.R-project.org/package=MBA

Many users and packages wish to display an image of a surface based on interpolation from possibly irregularly scattered points. Probably because an implementation of Akima's `akima::interp` function (1978) was provided in S-Plus, and developers and users familiar with `akima::interp` took up Albrecht Gebhardt's early port to R (on CRAN since 1998). Users of S-Plus were shielded from the ACM non-commercial use license because such issues were handled by their software provider. Users of the R **akima** package do however face the license conditions themselves, but have largely regarded the package as a useful adjunct to visualization ("for pretty smoothing" as put in a comment in the Bioconductor **OCplus** package), and so unproblematic in use.

In this contribution, we will assess the scale of the problem in terms of numbers of packages using **akima**, and establish which **akima** functions are used by these packages. Following a description of the foundations for 2D surface interpolation, the affected **akima** functions will be presented with possible alternatives from the **interp** package (under development) and the **MBA** package. Then follows an extensive simulation study using surfaces and point pattern shapes used in the literature. The study is intended to show which consequences may follow for package authours and maintainers switching from **akima** to unencumbered alternatives. We conclude that, in the vast majority of cases, the use of equivalent alternatives does not lead to any loss in functionality, and may lead to improvements in run times in many cases.

References

Akima, Hiroshi. 1978. "A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points." *ACM Transactions on Mathematical Software* 4: 148–59.

DeducerHansel: Econometrics in *R* through a Graphical User Interface

R. Scott Hacker[1]

1. Jönköping International Business School, Jönköping University, Scott.Hacker@ju.se

Keywords: GUI, **Deducer**, econometrics, time series, panel data

Webpages: https://rscotth.github.io/DeducerHansel/

This presentation discusses the Deducer plug-in, **DeducerHansel** (currently available through GitHub) that can deal with techniques typically found in undergraduate courses in econometrics, along with some more advanced econometric techniques. Currently the **Deducer** package (Fellows [1]) provides an exceptional interface that deals with a number of areas including generalized linear models. Thus it can already deal with ordinary least squares, weighted least squares, probit models and logit models. However it is not currently well-suited for dealing with time-series data, panel data, or censored data, or for dealing with instrumental variables. That is where **DeducerHansel** helps. The following areas are among those covered by Hansel: two-stage least squares; tobit models; smoothing, filtering, and forecasting; unit root testing; vector autoregressive models; cointegration testing; and various panel data and spatial data techniques. **DeducerHansel** can deal with the time series classes ts, zoo, and xts in addition to data frames. Hansel is similar in ease to the commercial software *EViews* and another open-source econometric software package called *gretl*, which is written in *C*. Hansel is not only useful for students in econometrics courses, but also provides an opportunity for those unacquainted with *R* to quickly get down to the business of using it for estimation. This can provide a gateway for deeper use of *R*.

References

[1] Fellows I (2012). Deducer: A Data Analysis GUI for *R*. Journal of Statistical Software 49(8), 1-15.

Assessing and visualizing drug synergy

Rytis Bagdžiūnas[1]

1. Open Analytics

Keywords: Synergy, Dose-Response, Response surface, Loewe additivity

Webpages: https://bigl.openanalytics.eu

Drugs are said to have synergistic effects if they increase each other's effectiveness when taken together. Combining such drugs can be beneficial in boosting treatment efficacy or reducing adverse effects to patient's health. **bigl** package aims to provide a generic, complete and intuitive workflow for assessing synergy effects in a two-drug study from a response surface perspective.

bigl workflow starts from response data on a grid of dose combinations and fits two 4-parameter log-logistic dose-response functions for each of the drugs by means of non-linear least squares or general purpose optimization methods. A flexible fitting procedure allows fixing or constraining parameters based on prior knowledge. Assessment of synergy effects can then be evaluated under 3 types of null models for expected response surface: generalized Loewe additivity (default), classical Loewe additivity or highest single agent. Formal statistical tests are applied to test for synergistic effects dataset-wise (`meanR`) or combination-wise (`maxR`). Bootstrapping can be used for estimating null distributions of these statistical tests and parallelization is available to speed up the process.

At the end of the procedure, user is presented with summary tables to evaluate global and combination-wise synergy effects. Color-coded 2-dimensional and rotatable 3-dimensional plots illustrating combination-wise synergy scores and other statistical quantities are also provided.

References

Koen Van Der Borght, Annelies Tourny, Rytis Bagdžiūnas, Olivier Thas, Maxim Nazarov, Heather Turner, Bie Verbist, and Hugo Ceulemans. 2017. "BIGL: Biochemically Intuitive Generalized Loewe Null Model for Prediction of the Expected Combined Effect Compatible with Partial Agonism and Antagonism."

William R. Greco, Gregory Bravo, and John C. Parsons. 1995. "The Search for Synergy: A Critical Review from a Response Surface Perspective." *Pharmacological Reviews* 47 (2). American Society for Pharmacology; Experimental Therapeutics: 331–85. http://pharmrev.aspetjournals.org/content/47/2/331.

causaleffect: An R Package for Causal Inference

Santtu Tikka[1] *and Juha Karvanen*[1]

1. University of Jyväskylä, Department of Mathematics and Statistics

Keywords: causal model, causal inference, graph, identifiability, causal effect

Webpages: https://CRAN.R-project.org/package=causaleffect

Causal models provide the formal framework for analyzing effects of external actions on a given causal system. The question of identifiability, i.e. whether the the interventional probability distribution can be uniquely computed using only observed probabilities has received substantial attention in literature. The *R* package **causaleffect** was created to provide an algorithm by (Shpitser and Pearl 2006) for computing causal effects (Tikka and Karvanen 2017).

Since its inception the **causaleffect** package has been significantly expanded. Often we have access to biased data only, and we wish to recover a causal effect of interest from the biased data. We may also have access to another population where experiments are possible, but some causal mechanisms may be different compared to the main population of interest. A recoverability algorithm by (Bareinboim and Tian 2015) and a transportability algorithm by (Bareinboim and Pearl 2013) are also provided by the **causaleffect** package to address these situations.

References

Bareinboim, Elias, and Judea Pearl. 2013. "Meta-Transportability of Causal Effects: A Formal Approach." In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (Aistats)*, 135–43.

Bareinboim, Elias, and Jin Tian. 2015. "Recovering Causal Effects from Selection Bias." In *Proceedings of the Twenty-Ninth Aaai Conference on Artificial Intelligence*, 3475–81. AAAI'15. AAAI Press.

Shpitser, Ilya, and Judea Pearl. 2006. "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." In *Proceedings of the 21st National Conference on Artificial Intelligence – Volume 2*, 1219–26. Boston: AAAI Press.

Tikka, Santtu, and Juha Karvanen. 2017. "Identifying Causal Effects with the R Package causaleffect." *Journal of Statistical Software* 76 (12): 1–30. doi:10.18637/jss.v076.i12.

eesim: Simulating Air Pollution a

Sarah Koehler[1,†] *and Brooke Anderson*[2]

1. Department of Statistics, Colorado State University
2. Department of Environmental and Radiological Health Sciences, Colorado State University
†Contact author: sakoehler7@gmail.com

Keywords: simulation, power analysis, epidemiology, environmental health

Simulation studies provide an important tool in developing epidemiologic methods to study the acute health risks associated with ambient exposures like air pollution and temperature. For example, studies have simulated time series of environmental exposure and health outcome data to compare time series and case-crossover methods; explore models to test for evidence of mortality displacement; investigate the impacts of measurement error and spatial misalignment of exposure and outcome measurements; and search for evidence of confounding by time-varying covariates. Developing the code to simulate environmental time series for such studies can be time consuming and simulation methods have varied across previous simulation studies, complicating the comparison of results across different studies. Here, we present eesim, an open-source R package that automates simulation of environmental time series of both exposure measurements and health outcomes with seasonal and long-term trends, offering both sensible defaults for each stage of the simulation process and also options to extensively customize specific simulation steps. Further, the software package provides tools for visualizing simulated time series and model performance results. This software provides efficient tools to environmental epidemiologists to explore the performance of environmental epidemiology models linking time series of health outcomes and ambient exposures, as well as to estimate the power of studies when planning or proposing future research.

Using R to protect athletes' health

Saulo Delfino Barboza[1] and Evert Verhagen[1,2,3]

1. Amsterdam Collaboration on Health & Safety in Sports, Department of Public & Occupational Health, VU University Medical Center Amsterdam, The Netherlands.
2. Australian Centre for Research into Injury in Sport and its Prevention, Federation University Australia.
3. UCT/MRC Research Unit for Exercise Science and Sports Medicine, Department of Human Biology, University of Cape Town, South Africa.

Keywords: epidemiology, surveillance, athletic injuries, sports medicine.

Webpages: https://safesports.amsterdam

Participation in sports exposes participants to an increased risk of injury and illness. At the recreational level, sports-related injury and illness result in societal costs, posing a burden for contemporary society. At the elite level, sub-optimal health and injury are detrimental for performance. Consequently, prevention is of great importance, and monitoring athletes' health is considered the first step towards effective prevention (Mechelen, Hlobil, and Kemper 1992). Therefore, our research group is interested in facilitating health surveillance in sports in order to early detect symptoms of injury and illness. Early detection enables early action to prevent minor health conditions to become bigger problems.

Basically, all the documentation of our sport surveillance system is organized in an *R* project in order to facilitate reproducible research. We have been using **mailR** to send a weekly online health registration questionnaire to athletes. Then we use **jsonline** to import, **tidyverse** to manipulate, and **rmarkdown** to report the data from athletes' to their respective team staff. With this strategy, we are able to capture athletes' health complaints and severity, and communicate on a weekly basis the outcomes to the responsible team staff. Then team staff follows up athletes individually and, when deemed necessary, act in order to prevent minor complaints to develop into time loss health issues.

We are also investigating the experiences of athletes and their staff (end-users) while using this surveillance system. We interviewed athletes and team staff, and coded the interview transcripts using **RQDA**. Team staff reported to be happy with the information in the weekly reports, but athletes wish more feedback and action on their data. As we are currently using **flexdashboards** and **plotly** to visualize data, we are investigating the possibility of generating personalized dashboards for athletes to visualize their own data and use it for their own purposes. A personalized dashboard can also be useful to provide tailored (evidence-based) feedback on health and safety behaviors. Although the possibility in theory exists (Verhagen and Bolling 2015), we are still investigating how to implement this in a feasible way.

References

Mechelen, Willem van, Hynek Hlobil, and Han C.G. Kemper. 1992. "Incidence, Severity, Aetiology and Prevention of Sports Injuries." *Sports Medicine* 14 (2). Springer Nature: 82–99. doi:10.2165/00007256-199214020-00002.

Verhagen, Evert, and Caroline Bolling. 2015. "Protecting the Health of the @Hlete: How Online Technology May Aid Our Common Goal to Prevent Injury and Illness in Sport." *British Journal of Sports Medicine* 49 (18). BMJ: 1174–8. doi:10.1136/bjsports-2014-094322.

Dynamic Item- and Teststatistics - A shiny GUI for test development

Sengewald Erik[1]

1. Federal Employment Agency, Germany

Keywords: shiny, psychometrics, test developent, reproducible research

The Federal Employment Agency in Germany provides a whole test system in order to consult their clients properly. The development of these tests follows a strict and standradized process. *R* and especially **shiny** allowes us to formalize the development process technically. The poster shows some of the features implemented in the corresponding tool. A main feature is the possibility to choose dynamically an item selection and thus define a test configuration. The tool shows some classical test parameter and allows for group and subsample specific analysis. Confirmatory Factor Analysis and SEM are applied, too. At the end it is possible to save the analysis together with the data, which ensures reproducibility and facilitate collaborative test development.

useR! 2017 Poster Presentation

Modelling High Frequency Financial Data

Smit Rohan

This paper studies non-linear behaviour of high frequency financial data and employs bivariate model for price change and duration (PCD model). There has been a growing interest in such models in the recent past to study the market micro-structure. The PCD model has been originally introduced by McCulloch and T say (2000). We have four simple conditional models to handle the dynamic structure within a trading day. We study the effect of number of transactions without price change on the size of price change. Price reversal is also an another important factor in this analysis wherein we study if the behaviour of the previous directions of the price change (whether price goes up or comes down) has a role in explaining the price reversal for stocks from different sectors. The order of dependence also changes with sector and hence problem of order selection is considered. Local trend in price movement has been studied by a general autoregressive structure. Instead of modelling durations by an ACD structure, we introduce additional explanatory variables in the model and investigate their significance. To estimate the parameters we have used Markov Chain Monte Carlo method for which we have used Gibbs sampling and Metropolis Hasting Algorithm. We estimated this model using high frequency data from NSE, India.

Optimized *R* implementation of collaborative filtering

Stefan Nikolic

SmartCat Company, Novi Sad

Keywords: recommender systems, collaborative filtering, machine learninig, optimization

Collaborative filtering (CF) is one of the most popular techniques for building recommender systems. It is based on extraction of relevant knowledge from existing user-item ratings, which help us make predictions about future ratings. One type of CF algorithms is memory-based (nearest neighbors method), where we use rating data to compute similarities between users or items. This can be used to predict future ratings, assuming that similar users/items will have similar ratings. In this presentation, we will show some of our recent work on improving the classic implementation of memory-based CF in *R*, thus making it applicable to larger datasets and significantly reducing the training time. Rating data is usually represented in a matrix format. For the optimization, we took advantage of matrix sparseness when calculating similarities and nearest neighbors, by utilizing appropriate R functions that efficiently deal with sparse data. This decreased the execution time by an order of magnitude comparing to a more classic implementation (e.g. **recommenderlab**). Also, we used partial matrix computations, which resulted in making our implementation applicable to large rating matrices on which classic implementation ran out of memory. The main idea was to divide large matrix into blocks and process them independently. This approach can decrease the execution time as well, considering that the blocks can be processed in parallel.

One by One: A Shiny Webapp for the Design and Analysis of Single-Case Experiments

Tamal Kumar De[1], Bart Michiels[1], Johan Vlaeyen[2] and Patrick Onghena[1]

1. Methodology of Educational Sciences Research Group, KU Leuven, Belgium
2. Health Psychology Research Group, KU Leuven, Belgium

Keywords: Single-case experiments, randomization test, shiny, GUI

Webpages: https://tamalkd.shinyapps.io/scda/, https://CRAN.R-project.org/package=SCRT, https://CRAN.R-project.org/package=SCVA, https://CRAN.R-project.org/package=SCMA

NA

A solution to this hindrance is to develop a web-based application which exposes functionality already coded in these packages in a user-friendly GUI. The platform chosen for that purpose is Shiny, an open source web application framework developed by RStudio. We first implemented the functions available in the above-mentioned packages in a carefully planned GUI. The GUI was split into tabs with a logical progression. In the first tab, the user can design the experiment by setting design parameters using the design functionality in the **SCRT** package. After the experiment has been conducted, the user can upload the observed data in the next tab. The third tab contains tools for visually analyzing the data using the **SCVA** package. In the fourth tab the user can perform randomization tests on the observed data using the **SCRT** package (Edgington and Onghena 2007). Finally, in the fifth tab there are further tools for meta-analysis and calculation of effect size measures from the **SCMA** package.

The biggest challenge in the process of development was making sure the GUI was easy enough for the average user to be able to use without any external help while also ensuring the GUI to be reactive enough to adapt and change options for parameters with every user input. Furthermore, we added numerous tips and error messages to alert and explain the correct format of inputs required and outputs saved to the user. We also added more functionality to the basic **SCRT**, **SCVA** and **SCMA** packages and adapted the webapp accordingly. A final challenge will be to host the webapp on RStudio or on the university servers and allowing external researchers access to use the webapp and provide valuable feedback which can then be incorporated as required.

References

Edgington, Eugene S., and Patrick Onghena. 2007. *Randomization Tests.* Fourth. Chapman & Hall/CRC.

Sustainable Development and Sustainable City

T. Davaakhuu[1] *and T. Navchaa*[2]

1. *School of Tourism and Land management, Mongolian National University,* *dvkhuu@gmail.com*
2. *School of Business, University of the Humanities*

Keywords: Sustainable city, Environment, Development, Ecology

Climate change, rapid economic and population growth, the sharp increase consumption and services, and resources depletion are the greatest global challenges that pose a risk for the earth's subsistence (1). Based on an estimation of ecological footprint, two Earth's will be required by 2030 as current consumption and production patterns continued (1). Sustainability has been widely debated in the construction industry in recent years. Through various assessments have previously been developed to help improving sustainability of city projects, those base networks for construction, transportation and usage of people. Therefore, Smart city/Green city is approach is most important others to becoming a sustainability. This study reveals to study for international green city index and after compare and analyzing Mongolian case before studying and implementation. Methodology processes begin with data and information collection by using international papers distribution to the expertise which involve in green city development and also green issues. After that, analyzing the Mongolian green city development and implementation. To find correct way of how to develop or to reach a smart city which problems have in Mongolian? Also to study what way is the best to improve it? After all, result can show Mongolian green city definitions which are waste management, transportation and construction are analyzed this paper.

R packages for decision making problems

T. González-Arteaga[2] *and R. de Andrés Calle*[2]

1. BORDA and PRESAD Research Groups, Multidisciplinary Institute of Enterprise (IME)
University of Valladolid, Spain
2. BORDA Research Unit, PRESAD Research Group, Multidisciplinary Institute of Enterprise (IME)
University of Salamanca, Spain

Keywords: Group decision making problems, Multi-Criteria Decision Making Methods, Decision Support, Outranking tools, Aggregation operator

Decision making problems usually refer to decisions by a group of experts that frequently involve multiple and conflicting criteria. In this regard, it is possible to find an extensive amount of contributions in the literature. Decision making problems and its study are a significant field nowadays, this fact produces an increasing number of developments and implementations of decision support tools. Although R has become very widespread among statisticians and data scientists, its use is not so much extended among researchers and users from Decision Making Theory. Since R is increasingly being used both in academic and non-academic environments, it is relevant to examine what R offers this field. The purpose of this presentation is showing how R could help decision making researchers and practitioners to implement their approaches. The poster brings to light a general overview of different R useful packages regarding this matter (some of them already available on CRAN). This contribution could be considered like a preliminary "CRAN task view".

Who infects whom? Generalised outbreak reconstruction using outbreaker2

Thibaut Jombart[1]
thibautjombart@gmail.com

1. Imperial College London.

The investigation of infectious disease epidemics increasingly relies on complex statistical analyses for reconstructing transmission chains ('who infects whom'), which can in turn yield precious insights into disease dynamics and inform containment strategies. These analyses typically rely on integrating various kinds of data, such as the dates of symptom onset, the location of the infected individuals, contacts between patients, and pathogen Whole Genome Sequences (WGS). In less than a decade, a variety of Bayesian methods have emerged to achieve such reconstruction, with few attempts at unifying different approaches under a single implementation framework. The R package *outbreaker2* aims to fill this gap. It provides a general framework for implementing Bayesian outbreak reconstruction using a modular approach, where data, priors, likelihoods, and MCMC algorithms all work as separate components. While each module is effectively implemented in C++ using Rcpp, *outbreaker2* allows for custom R functions to be specified by the user, so that various alternative models can be implemented seamlessly. *outbreaker2* also heavily relies on unit testing (*testhat*, *covr*) to facilitate continuous integration, and combines static (*ggplot2*) and interactive graphics (*htmlwidget*) to visualise results. After introducing the rationale of our approach, we will discuss some of the finer details of the implementation, the challenges between flexibility and computer efficiency, and the crucial role played by unit testing in the implementation of complex stochastic algorithms. We will conclude by discussing the potential role an open-source, modular statistical framework can play to shape the development of an emerging methodological field.

Developing R Tools for Health Risk for Ozone data

Tiantian Li[1]

1. National Institute of Environmental Health, Chinese Center for Disease Control and Prevention

Keywords: Ozone, Ozone pollution, raster data, zonal statistics, health risk

Ozone pollution is becoming increasingly severe in China which cause many health problem of people. In order to analyze future health risk under the RCP and SSP scenarios, research presented here applies data specialization, data extraction from global datasets, zonal statistics and field calculation on massive datasets. R can make the whole data processing more easily and elegantly. However, we found raster package, which is the most useful tools when handling with spatial data, has performance issue on large-scale raster processing. We also noticed gdalUtils package could do some similar operations, and its performance is great because of GDAL back end. But it is not convenience to set up GDAL in windows, and lack of zonal statistics function and some high level function such as crop and extract functions in raster package. It is therefore essential that we develop a useful package ROzone2 to solve these problems. It can automate install and setup for GDAL binary tools and warp some GDAL utilities, provide similar functions with raster package. And there is also parallel computation in ROzone2 to reduce time cost of a large amount of raster. Our poster demonstrates a suite of custom R packages to finish the process of Ozone data for analysis. Our current tools include: A package, ROzone2, which crop and rasterize raw data, and compute zonal statistics within the zones of a polygon in batch. A mapping app that allows user to visualize spatial distribution of the mean value of Ozone in the whole year. Features under development include: A Shiny app to help calculate health risk indicator (Ozone).

Parallelised Time Series Spike Detection using R on the Hadoop Platform

Timothy Wong[1], *Thomas Gerner*[2,6], *Kathy Chen*[3,6], *Caroline Alexander*[4,6], *Ondrej Urban*[5,6]

1. Centrica plc
2. Heidelberg University
3. Université de Nantes
4. National Aeronautics and Space Administration (NASA)
5. Stanford University
6. S2DS Pivigo

Keywords: time series, spike detection, Fourier transformation, MCMC, Hadoop

Smart meters records continuous stream of electricity consumption for each and every supply point across the United Kingdom. Energy suppliers are interested in understanding customer's consumption pattern in order to provide better service for them.

FlexiScore (F) is a new concept which British Gas has developed. It is a single numeric value ranging between 0 and 1 which quantifies the amount of flexible energy load for each electric supply points. High F value suggests the presence of erratic spikes, while low F value indicates prolonged consistency and non-spiky behaviour.

The algorithm has been productionised on the Hadoop platform (on premise) using Microsoft R Server 8.0 as a fully-scalable analytics framework. The large-scale distributed process contains an array of Markov Chains Monte Carlo (MCMC) for missing data permutation. A layer of Fourier transformation has been applied to create seasonal time series model. Afterwards, simple heuristics is applied to isolate erratic consumption spikes. The F score is then computed as output alongside other descriptive statistics.

What's in the network? A stepwise overview of working with networked data in *R*

Tine Van Calster[1], Michael Reusens[1], María Óskarsdóttir[1], Sandra Mitrović[1], Jasmien Lismont[1], Jochen De Weerdt[1], Wilfried Lemahieu[1], Bart Baesens[1], Jan Vanthienen[1]

1. KU Leuven, Dept. of Decision Sciences and Information Management, Naamsestraat 69, B-3000 Leuven, Belgium

Keywords: Featurization, Graphs, Networked Data, Social Network Analytics, Visualization

Networked data are inevitable in various real-life situations and domains. Basically, any collection of relationships between any arbitrary type of entities represents a network. Depending on different domains, these entities are typically called nodes and the links between them are referred to as edges. Concrete examples include: protein-protein interaction (PPI) networks in biology, transportation networks, social networks, retail networks (e.g. Amazon), citation networks, etc. This type of data is therefore used in many applications, ranging from fraud detection and churn prediction to the optimization of traffic. However, gaining insights from and fully exploiting the potential of networked data can be challenging. From gathering and structuring the data to building graphs and extracting information, the possibilities at each step in the process are abundant. In this abstract, we aim to provide an overview of this workflow as a whole, while discussing some of the possibilities available at each particular step and drawing special attention on *R* packages that can be utilized (independently or combined) for handling networked data.

Firstly, *structuring data* for efficient storing and manipulation is the initial step when working with networked data. The package **igraph** supports different structures that can be used for graph representations, e.g. adjacency matrices or edge lists. However, large graphs are often sparse, which requires some special attention. In *R*, the **Matrix** package can generate sparse matrices or the **slam** package can convert triplet representations into sparse matrices. Once the data structure is in order, we can take a look at *graph topologies*. We can distinguish between unipartite, bipartite, and n-partite graphs indicating the number of node types (e.g. authors and papers). Additionally, multigraphs are another type of graphs where the same pair of nodes can be connected with multiple (types of) edges. Due to the fact that real-life problems typically do not require capturing only network topology, but also different characteristics of nodes and/or edges, these are often enriched with additional attributes. This kind of networks are known as labeled networks. Depending on the topology and the final goal of our analysis, we can transform our graph, e.g. using the **Matrix** package to transform bipartite networks into unipartite ones, or add attributes, such as edge weights, in the **igraph** and **sna** packages.

Once the network is constructed, it can be used to gain new insights by different types of analysis. The first and most straightforward method is to simply *visualize* the graph, using **igraph**, **ggraph** or **sna**, to, for example, discover communities within the network. However, we can also extract *network features*, such as centrality measures (e.g. degree) that can be calculated using the **igraph** and **sna** packages, or features that can be derived from node/edge attributes. These network-based attributes can then play a vital role in e.g. classification applications. An igraph object with multiple node or edge attributes can easily be converted into a data.frame for further analysis. Thirdly, *network learning*, such as predicting links and labels of nodes, can be performed. Finally, the **igraph** package offers functions for graph sampling which can be useful for large networks.

Networked data can be complex and cumbersome to work with. In this abstract we presented an overview of the process and possibilities when working with networks. Only when tackled appropriately will the networks show us what they are really made of!

Applying Data Science For Social Good In Nonprofit Organization With Troubled Family Risk Profiling R Dashboard Application

Ting-Wei Lin[1],Tsun-Wei Tseng[2],Yu-Hsuan Lin[3],Pei-Yu Chen[4],Ning-Yuan Lyu[5],Ting-Kuang Lo[6],Shing-Yun Jung[7],Hsu Wei[8],Charles Chuang[9],Chun-Yu Yin[8],Johnson Hsieh[10,11]

[1]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica; [2]TAO Info Co.Ltd; [3]inQ Technology Co.Ltd; [4]Pegatron Co.Ltd; [5]Department of Electrical Engineering, National Tsing-Hua University; [6]Department of applied mathematics, Feng Chia University; [7]Institute of Computer Science and Engineering, National Chiao Tung University; [8]Hfoundation; [9]NETivism Co.Ltd; [1]0Department of Computer Science, National Chengchi University; [1]1DSP Co.Ltd

Keywords: troubled family risk profiling, data science for social good, shiny dashboard, integrated analysis

Webpages: Dashboard | GitHub

Assessing troubled families' risk status and distributing resources appropriately is a big issue for nonprofit organizations, not to mention national programs like the UK Troubled Families Program[1,2]. Given the increasing demand for social assistance and the long-term shortage of social workers, a more precise and continuous way to handle social resources wisely and efficiently is needed. Besides, junior social workers find it hard to quickly assess families through the many interview records and past documents, to decide whether to intervene. And most importantly of all, they lack the ability to utilize this information with proper data engineering and analysis. Here we apply an evidence-based approach on assessing the troubled families' risk status with a Shiny dashboard application integrating a prediction model generated from families' archives and follow-up records under the Data Science For Social Good Program in Taiwan, with the cooperation between a volunteer data science team and local nonprofit organization HFoundation.

To start, we use customer journey analysis to map the social worker's experience and organizational workflow. Overall, this covers 57 families accepting active aid, with a maximum follow-up time of 6 years. The documents from these families have basic socio-economic information and records from home visits with various follow-up time by the social workers. After de-identification of those documents, we pre-process the family archives and follow-up interviews into suitable formats for further management. Then, we use a risk factors system to tag each home visit in order to create a family risk prediction model. There are seven major risk factor categories ranging from financial problems to housing problems. We also use topic models to extract more information related to those risk factors. Then, we apply association rule analysis and discuss the resulting rules with the senior social workers. At the same time a steady state analysis with Markov chain was performed to calculate the recurrence rate of high risk factors for each home visit. These results can help to detect possible underlying risk factors and predict the possible recurrence rate for each risk factor from recent family status.

Based on the above analyses, a Shiny dashboard was built to assist social workers in their daily work. The dashboard provides a overview visualization of each case's family data in a timeline with risk factors and basic summary statistics. Social workers can easily get insights from past records and know the possible underlying risk factors with the association rules and decide which families's problems should be tackled first with the highest recurrence rate predicted from the model. In addition, the social workers can input their home visit data, which can update the model. In the end, the dashboard can be a great tool to provide nonprofit organizations, not only HFoundation, to precisely and efficiently manage their family cases.

References

1. Fletcher A, Gardner F, McKee M, Bonell C.(2012). The British government's Troubled Families Programme. BMJ2012;344:e3403. doi:10.1136/bmj.e3403
2. Chris Bonell, Martin McKee.(2016). Adam Fletcher. Troubled families, troubled policy making. BMJ 2016; 355 doi:10.1136/bmj.i5879

Frequency decomposition of connectedness measures

Tomáš Křehlík[1]

1. Charles University in Prague

Keywords: Connectedness measures, spectral domain, FEVD, multivariate systems

Webpages: https://github.com/tomaskrehlik/frequencyConnectedness

We develop and present a package that provides a frequency decomposition of the popular connectedness measure (Diebold and Yilmaz 2014) of a multivariate system. We present the main theoretical background as established in (Baruník and Křehlík 2017), practical implementation including the bootstrapping procedures, and present a case study. In the case study, based on (Baruník and Křehlík 2017), we decompose the systemic risk of the systemic risk of the main 13 financial institutions over the past 16 years providing an illustration of dynamism across frequencies.

References

Baruník, Jozef, and Tomáš Křehlík. 2017. "Measuring the Frequency Dynamics of Financial Connectedness and Systemic Risk."

Diebold, Francis X., and Kamil Yilmaz. 2014. "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms." *Journal of Econometrics* 182 (1). Elsevier: 119–34.

Interactive Shiny Applications for Flexible Modeling of Quantitative Predictors in Epidemiology

Crippa A.[1] *and Orsini N.*[1]

1. Deparment of Public Health Sciences, Karolinska Institutet

Keywords: Flexible Modeling, Dose-response Analysis, Shiny, Interactive Web Application, Regression Models, Epidemiology

Webpages: https://CRAN.R-project.org/package=mypkg, https://rpubs.com/username/project

Regression models (linear, logistic, Poisson, Cox, meta-analysis) are frequently used in epidemiology to investigate the dose-response relationships between quantitative predictors and the response variable. Flexible non-linear models involving transformations of the original predictors (e.g. quadratic, categories, splines terms) are common and widely used in medical research [1]. However, their implementation is not straightforward, as well as presentation of the corresponding results [2]. Therefore, we developed two interactive web applications in Shiny, flexmod[4] and dosresmeta[5], to facilitate estimation and presentation of flexible models. The shiny apps allow user to quickly explore different models without being familiar with any statistical software. They can be also useful for introducing the topic when teaching dose-response analysis. Our aim is to present the developed web applications to estimate several dose-response models using common epidemiological studies of both individual (case-control, cohort) and aggregated data (meta-analysis)[3]. The main focus will be on the interpretation and presentation of the findings either in a tabular or graphical form.

References

[1] Greenland S. Dose-response and trend analysis in epidemiology: altern- atives to categorical analysis. *Epidemiology.* 1995 Jul 1:356-65.

[2] Orsini, N. and Greenland, S., 2011. A procedure to tabulate and plot results after flexible modeling of a quantitative covariate. *Stata Journal*, 11(1), p.1.

[3] Orsini N, Li R, Wolk A, Khudyakov P, Spiegelman D. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *American Journal of Epidemiology.* 2012 Jan 1;175(1):66-73.

[4] Crippa A, Orsini N: Flexible modeling. [http://alessiocrippa.com/shiny/flexmod/]

[5] Crippa A, Orsini N: Performing dose-response meta-analysis. [http://alessiocrippa.com/shiny/dosresmeta/]

micemd: a smart multiple imputation R package for missing multilevel data

Vincent Audigier[1,2,3] and Matthieu Resche-Rigon[1,2,3]

1. Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, Paris
2. Université Paris Diderot - Paris 7, Sorbonne Paris Cité, UMR-S 1153, Paris
3. INSERM, UMR 1153, Equipe ECSTRA, Hôpital Saint-Louis, Paris
Contact author: vincent.audigier@univ-paris-diderot.fr

Keywords: Missing data; multiple imputation; multilevel data; individual patient data meta-analysis, systematically missing values

Statistical analysis often requires allowance for a multilevel structure. For example, a two-level structure occurs when individual data are aggregated, as in individual participant data (IPD) meta-analysis: individuals correspond to the lowest level, while the clinical study corresponds to the highest level. However, studies typically differ in their data collection and availability of confounders could varies. Consequently, by merging the studies, systematically missing data, i.e. missing for all individuals in a study, could be introduced. In addition, missing data can occur within each study (sporadically missing data). Unfortunately, statistical methods used for analysing such multilevel data cannot be straightforwardly applied on an incomplete data set. Multiple imputation is a common strategy to overcome the missing data issue. Several MI methods have been proposed in the literature to impute multilevel data with sporadically missing values only. Most of them use random effect models as imputation models. However, methods for dealing with more complex missing data, such as systematically missing variables, are needed. Recently, the issue has been addressed by several new approaches. Some multiple imputation methods for multilevel data have been implemented through several R packages (pan, jomo, mice, micemd). However, they are not all tailored for handling sporadically and systematically missing values, or for continuous and binary variables. The main objective of this talk is to present the new package micemd, which is based on a new methodology (two stage estimators), and to provide some guidelines for using the suitable package according to the dataset which is analysed. First, the talk is motivated by an IPD meta-analysis in cardiovascular disease consisting of 28 observational cohorts in which systematically missing and sporadically missing data occur (GREAT data). Then, based on a simulation study, the advantages and drawbacks of each multiple imputation method are discussed. Finally, the multiple imputation methods are applied to the GREAT data.

Model-based detection of differences in dynamic systems

Wolfgang Mader[1], Marcus Rosenblatt[1], Mirjam Fehling-Kaschek[1], Jens Timmer[1], and Daniel Kaschek[1]

1. *University of Freiburg, Germany*

Keywords: Systems Biology, Parameter Estimation, Regularization

Webpages: https://github.com/dkaschek/dMod, https://github.com/dkaschek/cOde

Systems biology is an interdisciplinary field with the aim of understanding biological systems on the cellular level based on mathematical models. Using time-resolved measurements of interacting states, e.g., molecular species in cell signaling pathways, a mechanistic understanding of the cell functioning can be gained. In particular, ordinary differential equations (ODE) are used to model the dynamics of interacting states. The parameters of the ODEs characterize the realized behaviour of a cell or a cell population. Therefore, a fundamental step is to fit the parameters of the model to experimental data and to construct confidence bounds for parameter estimates. To this end, the *R* package **dMod** is developed in our group, which together with **cOde** provides a convenient and fast modeling environment for dynamic systems modeling.

In systems biology of cancer or disease progression, one is regularly interested in the difference between a healthy and a diseased cell. By first describing the different cell types by ODEs and then finding the minimal differing set of parameters which allows to fit the data of all cell types, the differences in the cell behaviour are detected in a model-based manner.

One way to find this minimal set is L1 penalization during parameter estimation. We discuss the L1 approach together with several others and present a simulation study evaluating the performance of the different approaches.

useR! 2017 Poster Presentation

Development of R/shiny applications for the biopharmaceutical industry

Xavier Lories

17 mars 2017

The drug development and production processes generally requires a tremendous amount of statistical supports, both in the clinical and the non-clinical areas.

In some cases, this support is very specific and require the full attention of a statistician. In other cases, the support is required for very repetitive analyses, for which the statistical methodology to be applied is well-defined. In the latter case, the development of applications to automate the analysis may prove itself the best solution to optimize time and resource allocation.

R/shiny offers a convenient tool for the development of tailored applications. In this session, example of such applications, developed for clinical and non-clinical purposes will be presented with their objectives.

In regulated environments, these types of solutions often require a software validation (e.g. GAMP) to ensure the application is fit-for-purpose and deliver reproducible and high quality results. Unless carefully planned, this validation phase can become a very lengthy and costly step. Emphasis will be set on the validation of applications involving Bayesian models, which, due to their simulation nature, may be hard to validate.

Predict Hospital Length of Stay with R

Xinwei Xue[1], *Carl Saroufim*[1]
1. Data Group, Microsoft

Keywords: Hospital Length of Stay Prediction, R, Microsoft R Server, SQL Server

Webpages: https://microsoft.github.io/r-server-hospital-length-of-stay/

Abstract:

Hospital Length of Stay (LOS) is defined in number of days from the initial admit date to the date that the patient is discharged from any given hospital facility. There can be significant variation of LOS across various facilities and across disease conditions and specialties even within the same healthcare system. Accurately predict the length of stay of a patient upon admission is very important to hospitals. Advanced accurate LOS prediction at the time of patient admission will enable proper planning of resources, managing operational efficiency, reducing readmission risk and enhancing the quality of care and patient satisfaction.

In this talk, We'll talk about how to build an end-to-end solution to predict hospital LOS with *R* and Microsoft R Server. We'll talk about the sample data and model development steps (data processing, feature engineering, model training, scoring and evaluation) with R and SQL, and discuss the modeling results. Details can be found in github here

SQL Server 2016 provides R Services that allows in-database advanced analytics with R and Microsoft R Server (which eliminates the memory constraint and enables scalable computing when processing large data). This service allows computing to happen at the database server without moving the data out of the database, which mitigates risks associated with data movement. We'll show how to deploy R solutions as SQL stored procedures, which can be called from any applications or devices.

Finally, we show visualize the KPIs with Power BI Dashboard.

In the end we'll show the users how to deploy this solution to Azure Data Science Virtual Machine with one click.

References

Hospital LOS Prediction in Cortana Intelligence Gallery

SQL Server R Services

Microsoft R Server

DDIR : handling social research data standard with R

Yasuto NAKANO[1]

1. Kwansei Gakuin University

Keywords: DDI, xml, data format, social research data, reproducible research

Webpages: http://www.soc-nakano.net/?DDIR

The purpose of this presentation is to propose an useful tool for social research data and its analysis. **DDIR** is an *R* package which handles informations described in the DDI standard on *R* environment (NAKANO 2015; NAKANO 2016).

The Data Documentation Initiative (*DDI*) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences (DDI Alliance 2017). *DDI* is mainly implemented in large data archives and huge research projects to ensure their data handling consistent.

On the other hand, even small research projects or individual researchers could benefit from *DDI.* Because a *DDI* file collects/contains all informations we need in research activities (e.g. research questions, variable conceptualizations, questionnaire sentenses, variable names, value labels, etc.), it is efficient to use one *DDI* file as the sole source of informations in any steps of research activities.

In *R* environment, there is no standard data format for social research data. In many case, we have to prepare numerical data and label or factor informations separately. If we use *DDI* file as a data file with **DDIR** in *R*, only one *DDI* file is needed to be prepared. *DDI* could be a standard data format of social research data in *R* environment, just same as 'sav' file in *SPSS*.

We can retrieve necessary informations from a *DDI* file with **DDIR**. Further more, we can integrate and export informations related to the data as an *DDI* file with **DDIR**. The **DDIR** package realizes integrated social research analysis environment with *R*, and ensures it as a reproducible research.

References

DDI Alliance. 2017. http://www.ddialliance.org/.

NAKANO, Yasuto. 2015. "DDIR and Dlcm : Integrated Environment for Social Research Data Analysis." In *UseR!2015: R User Conference 2015*. Aalborg, Denmark.

———. 2016. "DDIR: An R Package for Handling Ddi Files." In *EDDI2016: Annual European Ddi User Conference*. Cologne, Germany.

Development of Data Base Integrated Hydrological- and Hydraulic Modeling for River Flood- and Urban Inundation Forecast

Yoshihiro Shibuo[1],Hiroshi Sanuki[2], SungAe Lee[3], Kohei Yoshimura[4], Yoshimitsu Tajima[5], Hiroaki Furumai[3], and Shinji Sato[5]

1. Int. Centre for Water Hazard and Risk Management, Public Works Research Institute
2. Penta-Ocean Construction Institute of Technology
3. Research Center for Water Environment Technology, The University of Tokyo
4. Research Organization for Regional Alliances, Kochi University of Technology
5. Department of Civil Engineering, The University of Tokyo

Keywords: River flood, Urban inundation, Model forecast, Hydrology, Hydraulics

There has been increasing demand in forecasting inundation in coastal urbanized areas, which is vulnerable to river flood from upstream, urban inundation due to torrential rainfall, or storm surges from coast. In order for realizing the forecast in such areas, integrated modelling framework of river flood, urban sewer network, coastal hydraulics is necessary for seamlessly exchanging model input and output, such as water levels, and discharges, among models. It is also crucial that the model is effectively integrated into a database of required forcing so that early warnings shall be issued on time. The present research demonstrates an ongoing development of such a seamlessly integrated model and the wrapping system that connects the model to an earth- and environmental data archives for feeding required forcing to the model. Being linked to the data base online, the system dynamically handles miscellaneous spatiotemporal data, e.g., ground radar observed rainfall, water levels in river outlets etc., using GIS functions in R environment, and drive the seamless model. The system employs R functions also for visualization of moving rainfall, inundated areas, and underground state of sewer network during flood events. This modeling framework is intended to promote communications among relevant authorities of river management office, urban sewer management office, and coastal management office so that impact of water related hazard can be best minimized by integrated and effective flood counter measures.