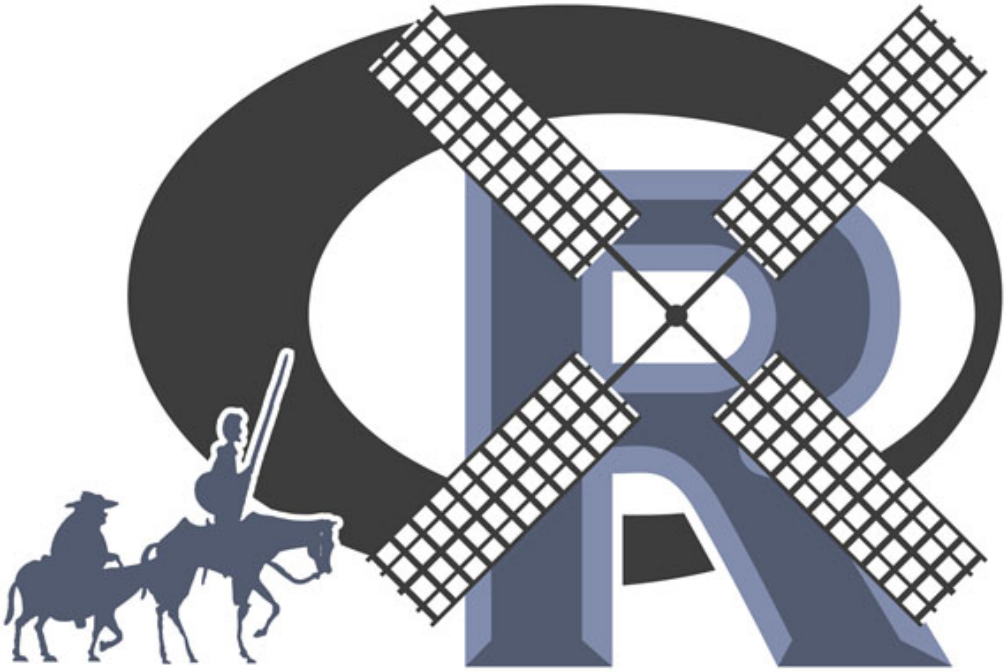
The R User Conference, useR! 2013
July 10-12 2013
University of Castilla-La Mancha, Albacete, Spain

Book of Contributed Abstracts

Compiled 2013-07-01



Contents

Wednesday 10th July 6

Bioinformatics, 10:30 **6**
- Integrating R with a Platform as a Service cloud computing platform for Bioinformatics applications. . 7
- Simulation of molecular regulatory networks with graphical models 8
- GOsummaries: an R package for showing Gene Ontology enrichment results in the context of experimental data 9
- Analysis of qPCR data in R 10

Computational Challenges in Molecular Biology I, 10:30 **11**
- The GenABEL suite for genome-wide association analyses 11
- Making enzymes with R 12
- Use of molecular markers to estimate genomic relationships and marker effects: computation strategies in R 13
- High Content Screening Analysis in R 14

Environmental statistics I, 10:30 **15**
- rClr package - low level access to .NET code from R 15
- Reproducible Research in Ecology with R: distribution of threatened mammals in Equatorial Guinea. . 16
- Using R for Mapping the Spatial Extent of Meteorological and Hydrological Drought Events 17

Statistics/Biostatistics I, 10:30 **18**
- Three-component decomposition of coal spectrum in R 18
- Method of comparison of actions of the liquidators of the accident on Chernobyl Nuclear Power Plant on the basis of fragmentation of their routes and encryption it in a form similar to the DNA 19
- Differential expression analysis of RNA-seq data at base-pair resolution in multiple biological replicates 20
- Statistical inference for Hardy-Weinberg equilibrium with missing data 21

Computational Challenges in Molecular Biology II, 12:20 **22**
- What did we learn from the IMPROVER Diagnostic Signature Challenge? 22
- Deciphering the tRNA operational code - using R 23
- Big Data and Reproducibility – Building the Bridge 24
- Topology-based Hypothesis Generation on Causal Biological Networks using igraph 25

Econometric Software, 12:20 **26**
- Hansel: A Deducer Plug-In for Econometrics 26
- Robust standard errors for panel data: a general framework 27
- Rsiopred: An R package for forecasting by exponential smoothing with model selection by a fuzzy multicriteria approach 28
- AutoSEARCH: Automated General-to-Specific Model Selection 29

Environmental statistics II, 12:20 **30**
- Driving R to the air quality industry. NanoEnvi Analyst: a tool for designing large-scale air quality plans for improvement in ambient air quality 30
- Sequential Design of Experiments for model selection: an application to the energy sector 31
- Emission inventory supported by R: dependency between calorific value and carbon content for lignite . 32

Statistics/Biostatistics II, 12:20 **33**
- Leveraging GPU libraries for efficient computation of Gaussian process models in R 33
- TriMatch: An R Package for Propensity Score Matching of Non-Binary Treatments 34
- KmL3D: K-means for Joint Trajectories 35
- Stochastic Modeling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa Disrict 36

Database applications, 16:30 **37**
- Introducing SimpleDataManager - A simple data management workflow for R 37
- SenseLabOnline: Combining agile data base administration with strong data analysis 38
- ffbase: statistical functions for large datasets 39

Statistics/Biostatistics III, 16:30 **40**
cold: a package for Count Longitudinal Data . . . 40
kPop: An R package for the interval estimation of the mean of the selected populations. . . . 41
GLM - a case study: Antagonistic relationships between fungi and nematodes . . . 42
R Packages for Rank-based Estimates . . . 43

Time Series Analysis, 16:30 **44**
Heart Rate Variability analysis in R with RHRV . . . 44
Massively Parallel Computation of Climate Extremes Indices using R . . . 45
Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data . . . 46
hts: R tools for hierarchical time series . . . 47

Using R for Teaching I, 16:30 **48**
Teaching statistics interactively with Geogebra and R . . . 48
RKTeaching: a new R package for teaching Statistics . . . 49
genertest: a package for the developing exams in R . . . 50
Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond . . . 51
Teaching R in the Cloud . . . 52

Thursday 11th July 53

Machine learning I, 10:00 **53**
BayesClass: An R package for learning Bayesian network classifiers . . . 53
Constructing fuzzy rule-based systems with the R package "frbs" . . . 54
bbRVM: an R package for Ensemble Classification Approaches of Relevance Vector Machines . . . 55
Classification Using C5.0 . . . 56

Marketing/Business Analytics I, 10:00 **57**
Extending the Reach of R to the Enterprise . . . 57
Big-data, real-time R? Yes, you can. . . . 58
Large-Scale Predictive Modeling with R and Apache Hive: from Modeling to Production . . . 59
Non-Life Insurance Pricing using R . . . 60

Official statistics I, 10:00 **61**
ReGenesees: symbolic computation for calibration and variance estimation . . . 61
Big data exploration with tabplot . . . 62
rwiot: An R package for Input-Output analysis on the World Input Output Database (WIOD) . . . 63
Make Your Data Confidential with the sdcMicro and sdcMicroGUI packages . . . 64

Statistical Modelling I, 10:00 **65**
MRCV: A Package for Analyzing the Association Among Categorical Variables with Multiple Response Options . . . 65
Different tests on lmer objects (of the lme4 package): introducing the lmerTest package. . . . 66
Implementation of advanced polynomial chaos expansion in R for uncertainty quantification and sensitivity analysis . . . 67
Dhglm & frailtyHL : R package for double hierarchical genralized linear models and frailty models . . . 68

Machine learning II, 11:50 **69**
rknn: an R Package for Parallel Random KNN Classification with Variable Selection . . . 69
Patterns of Multimorbidity: Graphical Models and Statistical Learning . . . 70
ExactSampling: risk evaluation using exact resampling methods for the k Nearest Neighbor algorithm . . . 71
Classifying High-Dimensional Data with the The HiDimDA package . . . 72

Marketing/Business Analytics II, 11:50 **73**
Groupon Impact Report: Using R To Power Large-Scale Business Analytics . . . 73
Statistics with Big Data: Beyond the Hype . . . 74
Using survival analysis for marketing attribution (with a big data case study) . . . 75
Big Data Analytics - Scaling R to Enterprise Data . . . 76

Official statistics II, 11:50 **77**
Using R for exploring sampling designs at Statistics Norway . . . 77
Application of R in Crime Data Analysis . . . 78
Maps can be rubbish for visualising global data : a look at other options. . . . 79
The use of demography package for population forecasting . . . 80

Statistical Modelling II, 11:50 **81**
Shape constrained additive modelling in R . . . 81
Semiparametric bivariate probit models in R: the SemiParBIVProbit package . . . 82
"RobExtremes": Robust Extreme Value Statistics — a New Member in the RobASt-Family of R Packages 83
Generalized Bradley-Terry Modelling of Football Results . . . 84

Biostatistics: Regression Methodology, 16:30 **85**
Copula sample selection modelling using the R package SemiParSampleSel . . . 85
Robust model selection for high-dimensional data with the R package robustHD . . . 86
HGLMMM and JHGLM: Package and codes for (joint)hierarchical generalized linear models . . . 87
Fitting regression models for polytomous data in R . . . 88

Programming, 16:30 **89**
An exposé of naming conventions in R . . . 89
Statistical Machine Translation tools in R . . . 90
Reference classes: a case study with the poweRlaw package . . . 91
Combining R and Python for scientific computing . . . 92

R in companies, 16:30 **93**
Shiny: Easy web applications in R . . . 93
rapport, an R report template system . . . 94
Seamless C++ Integration with Rcpp Attributes . . . 95
The R Service Bus: New and Noteworthy . . . 96

R in the Central Banks, 16:30 **97**
Outliers in multivariate incomplete survey data . . . 97
Use of R and LaTeX for periodical statistical publications . . . 98
Solving Dynamic Macroeconomic Models with R . . . 99

Kaleidoscope I, 18:20 **100**
packdep: network abstractions of CRAN and Bioconductor . . . 100
The Beatles Genome Project: Cluster Analysis of Popular Music in R . . . 101
The secrets of inverse brogramming . . . 102

Kaleidoscope II, 18:20 **103**
Mapping Hurricane Sandy Damage in New York City . . . 103
Unlocking a national adult cardiac surgery audit registry with R . . . 104
Renjin: A new R interpreter built on the JVM . . . 105

Friday 12th July 106

GUIs/Interfaces, 10:00 **106**
Using Lazy–Evaluation to build the G.U.I. . . . 106
Survo for R - Interface for Creative Processing of Text and Numerical Data . . . 107
Using R in teaching statistics, quality improvement and intelligent decision support at Kielce University of Technology . . . 108

High performance computing, 10:00 **109**
Facilitating genetic map construction at large scales in R . . . 109
Elevating R to Supercomputers . . . 110
R in Java: Why and How? . . . 111
Rhpc: A package for High-Performance Computing . . . 112

Modelling and Optimization, 10:00 **113**
DCchoice: a package for analyzing dichotomous choice contingent valuation data . . . 113
Systems biology: modeling network dynamics in R . . . 114
Evolutionary multi-objective optimization with R . . . 115
An integrated Solver Manager: using R and Python for energy systems optimization . . . 116

Visualization/Graphics I, 10:00 **117**
Radar data acquisition, analysis and visualization using reproducible research with Sweave . . . 117
Network Visualizations of Statistical Relationships and Structural Equation Models . . . 118
tableR - An R based approach for creating table reports from surveys . . . 119
likert: An R Package for Visualizing and Analyzing Likert-Based Items . . . 120
Design of likert graphics with lattice and mosaic . . . 121

High performance computing II, 11:50 **122**
Open Source Product Creation, Bosco Team . . . 122
Practical computer experiments in R . . . 123
Symbiosis - Column Stores and R Statistics . . . 124
Memory Management in the TIBCO Enterprise Runtime for R (TERR) . . . 125

Reproducible Research, 11:50 **126**
TiddlyWikiR: an R package for dynamic report writing. . . . 126
Synthesis of Research Findings Using R . . . 127
compreGroups updated: version 2.0 . . . 128

Statistical Modelling III, 11:50 **129**
BayesVarSel. An R package for Bayesian Variable Selection. . . . 129
Bayesian learning of model parameters given matrix-valued information, using a new matrix-variate Gaussian Process. . . . 130
FluDetWeb: an interactive web-based system for the early detection of the onset of influenza epidemics 131
Looking for (and finding!) hidden additivity in complete block designs with the hiddenf package. . . . 132

Visualization/Graphics II, 11:50 **133**
A ggplot2 builder for Eclipse/StatET and Architect . . . 133
Visualizing Multivariate Contrasts . . . 134
metaplot: Flexible Specification for Forest Plots . . . 135
GaRGoyLE: A map composer using GRASS, R, GMT and Latex . . . 136

Regular Posters **137**
Asymmetric Volatility Transmission in Airline Related Companies in Stock Markets . . . 137
A R tool to teach descriptive statistics . . . 138
Using R to estimate parameters from multiple frames . . . 139
Calibration in Complex Survey using R . . . 140
R/Statistica Interface . . . 141
AMOEBA+ with R . . . 142
Software developments for non-parametric ROC regression analysis . . . 143
An R-package forWeighted Smooth . . . 144
Using R as continuous learning support in Sea Sciences degree . . . 145
Variable selection algorithm implemented in FWDselect . . . 146
Panel time series methods in R . . . 147
Teaching introductory statistics to students in economics: a comparison between R and spreadsheet . . 148
TestR: R language test driven specification . . . 149
Small area data visualization using ggplot2 library . . . 150
R as a Data Operating System for the Cloud . . . 151
TPmsm: Estimation of the Transition Probabilities in 3-State Models . . . 152
Climate Analysis Tools - An operational environment for climate products . . . 153
seq2R: Detecting DNA compositional change points . . . 154
NPRegfast: Inference methods in regression models including factor-by-curve interaction . . . 155
Pharmaceutical market analysis with R . . . 156
Standardisation on Statistics: ISO Standards and R Tools . . . 157
Quantitative Text Analysis of readers' contributions on Japanese daily newspapers . . . 158
Analysis of data from student surveys at Kielce University of Technology using R Commander and R Data Miner . . . 159

Statistical analysis with R of an effect of the air entrainment and the cement type on fresh mortar properties . . . 160
gxTools: Multiple approaches integrated in automated transcriptome analysis . . . 161
A cloud infrastructure for R reports . . . 162
On thinning spatial polygons . . . 163
Statistical analysis in R of environmental and traffic noise in Kielce . . . 164
Using R for dosimetry extremum tasks . . . 165
Data mining with Rattle . . . 166
intRegGOF: Modelling with the aid of Integrated Regression Goodness of Fit tests. . . . 167
An R script to model monthly climatic variables with GLM to be used in hydrological modelling . . . 168
Using R2wd package to automatize your reporting from R to Microsoft Word document - An application of automatic report for a survey in telecommunication . . . 169
Automation of spectroscopic data processing in routine tests of coals using R . . . 170
A Web-based Application as a Dynamical Tool for Clinical Trial Researchers . . . 171
Analysis of load capacity of pipes with CIPP liners using R Rattle package . . . 172
Efficiency analysis of companies using DEA model with R . . . 173
Introducing statistic and probability concepts with R in engineering grades . . . 174
Biomarker Discovery using Metabolite Profiling Data: Discussion of different Statistical Approaches. . . . 175
edeR: Email Data Extraction using R . . . 176
Reproducible and Standardized Statistical Analyses using R . . . 177
hwriterPlus: Extending the hwriter Package . . . 178
Application of the nearest neighbour indices in spatstat R package for Persian oak (Quercus brantii var. persica) ecological studies in Zagros woodlands, Iran . . . 179
Point process spatio-temporal product density estimation with R . . . 180
Spatio-Temporal ANOVA for replicated point patterns using R . . . 181
Estimation of parameters using several regression tools in sewage sludge by NIRS . . . 182
Recipe for the implementation of a population dynamics bayesian model for anchovy: Supercomputing using doMC , rjags and coda R packages . . . 183

Integrating R with a Platform as a Service cloud computing platform for Bioinformatics applications

Hugh P. Shanahan[1*], **Anne M Owen**[2], **Andrew P. Harrison** [2,3]

1. Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, U.K.,
2. Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, U.K.
3. Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, U.K.

*Contact author: hugh.shanahan@rhul.ac.uk

Keywords: Cloud Computing, GeneChips, Azure, PaaS, Microarray

Cloud Computing is increasingly being used by Bioinformatics researchers as well as by the scientific community in general. This has been largely encouraged by the rapid increase in the size of Omic data sets Stein (2010). There are advantages in using a cloud for short usages of powerful computers when scaling up programs which have been tested on a small amount of data. Much of the emphasis has been on the use of Infrastructure as a Service platforms, such as Amazon's EC2 service where the user gets direct access to the console of the Virtual Machines(VM's) and *MapReduce* frameworks, in particular *Hadoop* Taylor (2010). An alternative to this is to use a Platform as a Service (PaaS) infrastructure, where access to the VM's is programmatic. An example of this is the Microsoft Azure platform which we have made use of via the VENUS-C EU network.

A PaaS interface can offer certain advantages over the other approaches. In particular, it is more straightforward to design interfaces to software packages such as *R* and it obviates the need to port codes designed for single processors into a *MapReduce* framework. In the case of Azure, another advantage is that Microsoft Research have provided a set of *C#* libraries called the Generic Worker which allow easy scaling of VM's.

We have developed software that makes use of these libraries to run *R* scripts to analyse almost all of a specific microarray data set (HG_U133A - an Affymetrix GeneChip for humans) in the public database ArrayExpress. We have previously demonstrated that a small set of publicly deposited experiments that use this type of microarray are susceptible to a bias due to specific sequences that probes of the microarray hybridise with (runs of 4 or Guanines) Shanahan et al. (2011). We have used Azure to extend our analysis to 576 experiments deposited at ArrayExpress before May, 2012. In particular we have shown that correlations between probe sets can be significantly biased, suggesting that probe sets that have such probes will be more correlated with each other than they should be. This will bias a large number of conclusions that have been drawn on the basis of individual experiments and conclusions based on the inference of gene networks using correlations between probe sets over many experiments.

This analysis provides an exemplar to run multiple *R* jobs in parallel with each other on the Azure platform and to make use of its mass storage facilities. We will discuss an early generalisation we have dubbed **GWydiR** to run any *R* script on Azure in this fashion, with a goal on providing as simple a method as possible for a user to scale up their *R* jobs.

References

Shanahan, H. P., Memon, F. N., Upton, G. J. G. and Harrison, A. P. (2011, December). Normalized Affymetrix expression data are biased by G-quadruplex formation. *Nucleic Acids Research 40*(8), 3307-3315.

Stein, L. D. (2010, January). The case for cloud computing in genome informatics. *Genome biology 11*(5), 207.

Taylor, R. C. (2010, January). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics 11 Suppl 1*, S1.

Simulation of molecular regulatory networks with graphical models

Inma Tur[1], **Alberto Roverato**[2], **Robert Castelo**[1,⋆]

1. Universitat Pompeu Fabra
2. Università di Bologna
⋆Contact author: robert.castelo@upf.edu

Keywords: Molecular regulatory network, Graphical model, Covariance matrix, Simulation

High-throughput genomics technologies in molecular biology produce high-dimensional data sets of continuous and discrete readouts of molecules within the cell. A sensible way to scratch at the underlying complex network of regulatory mechanisms using those data is to try to estimate the graph structure G of a graphical model (Lauritzen, 1996). A fundamental step taken by many of the contributions to this problem is to test first the performance of the proposed algorithms on data simulated from a graphical model with a given graph G, before showing the merits of the approach on real biological data.

Here we introduce the functionality available in the R/Bioconductor package **qpgraph** (Tur et al., 2013) to simulate Gaussian graphical models, homogeneous mixed graphical models and data from them. The former produce multivariate normal observations which can be employed to test algorithms inferring networks from gene expression data, while the latter produce mixed discrete and continuous Gaussian observations, which can be employed to test algorithms inferring networks from genetical genomics data produced by genotyping DNA and profiling gene expression on the same biological samples.

A basic component to this functionality is the generation of a covariance matrix Σ with: (1) a pattern of zeroes in its inverse Σ^{-1} that matches a given undirected graph $G = (V, E)$ on $p = |V|$ vertices associated to $X_1, \ldots, X_p$ continuous Gaussian random variables; and (2) a given mean marginal correlation ρ for those pairs of variables connected in G. This is achieved by applying a matrix completion algorithm (Hastie et al., 2009, pg. 634) on a $p \times p$ positive definite matrix drawn from a Wishart distribution whose expected value is determined by ρ with $-1/(p-1) < \rho < 1$ (Odell and Feiveson, 1966). Building up on this feature, the package can interpret this matrix as a conditional one $\Sigma \equiv \Sigma(i)$, given a probability distribution on all joint discrete levels $i \in \mathcal{I}$, and simulate conditional mean vectors $\mu(i)$ with given linear additive effects, which enable simulating homogeneous mixed graphical models. Using the **mvtnorm** package, conditional Gaussian observations are simulated accordingly. This functionality is also integrated with the one of the **qtl** package for generating genotype data from experimental crosses to enable the simulation of genetical genomics data under some of the genetic models available in **qtl**. Critical parts of the code are implemented in *C* language enabling the efficient simulation of graphical models involving hundreds of random variables.

The technical complexity behind all these features is hidden to the user by means of S4 classes and methods that facilitate the simulation of these data, as illustrated in the vignette included in the **qpgraph** package (Tur et al., 2013) and entitled "Simulating molecular regulatory networks using qpgraph".

References

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning*. Springer.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press.

Odell, P. and A. Feiveson (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association 61*(313), 199–203.

Tur, I., A. Roverato, and R. Castelo (2013). The **qpgraph** package version 1.16.0. http://www.bioconductor.org/packages/release/bioc/html/qpgraph.html.

GOsummaries: an R package for showing Gene Ontology enrichment results in the context of experimental data

Raivo Kolde[1,2,*], **Jaak Vilo**[1,2]

1. Institute of Computer Science, University of Tartu, Liivi 2- 314, 50409 Tartu, Estonia
2. Quretec, Ülikooli 6a, 51003 Tartu, Estonia
*Contact author: rkolde@gmail.com

Keywords: principal component analysis, word clouds

Gene Ontology (GO) enrichment analysis is a common step in analysis pipelines for large genomic datasets. With the help of various visualisation tools, the interpretation of the enrichment results is rather straightforward, when the number of queries is small. However, as the number of queries grows the tools become less effective and it gets harder to gain a good overview of results. We introduce a novel R package **GOsummaries** that visualises the GO enrichment results as concise word clouds. These word clouds can be combined together into one plot in case of multiple queries. By adding also the graphs of corresponding raw experimental data, **GOsummaries** can create informative summary plots for various analyses such as differential expression or clustering. This approach is particularly effective for Principal Component Analysis (PCA). It is possible to annotate the components using GO enrichment analysis and display this information next to the projections to the components. The **GOsummaries** package is available at GitHub (https://github.com/raivokolde/GOsummaries)

Analysis of qPCR data in R

Laure Cougnaud[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: laure.cougnaud@openanalytics.eu

Keywords: qPCR, RT-qPCR, data analysis

qPCR is nowadays a standard procedure to quantify DNA copies with precision. However, as this technique is more and more used, criteria and consensus on methods to analyze qPCR data are lacking (Bustin et˜al., 2009). At the mean time *R* packages to analyze such data have flourished during the last decade.

In this presentation, we will present the main issues encountered when analyzing qPCR data, from the pre-processing (filtering, quality check of the experiment) to the statistical analysis itself. Then we will provide an overview of current *R* packages (**SLqPCR**, **EasyqpcR**, **ddCt**, **NormqPCR**, **ReadqPCR**,...) to analyze qPCR data and will assess their strengths and weaknesses in dealing with these issues.

References

Andersen, C.˜L., J.˜L. Jensen, and T.˜F. Ørntoft (2004). Normalization of real-time quantitative reverse transcription-pcr data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research˜64*(15), 5245–5250.

Bustin, A. and T.˜Nolan (2004). Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomol Tech˜15*(3), 155–166.

Bustin, S.˜A., V.˜Benes, J.˜A. Garson, J.˜Hellemans, J.˜Huggett, M.˜Kubista, R.˜Mueller, T.˜Nolan, M.˜W. Pfaffl, G.˜L. Shipley, J.˜Vandesompele, and C.˜T. Wittwer (2009). The MIQE guidelines : Minimum information for publication of quantitative real-time pcr experiments. *Clinical Chemistry˜55(4)*.

Dorak, M.˜T. (2006). *Real-Time PCR (Advanced Methods Series)*. Oxford: Taylor & Francis.

Perkins, J., J.˜Dawes, S.˜McMahon, D.˜Bennett, C.˜Orengo, and M.˜Kohl (2012). ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (cq) data. *BMC Genomics˜13*(1), 296.

Rieu, I. and S.˜J. Powers (2009). Real-time quantitative RT-PCR: Design, calculations, and statistics. *The Plant Cell˜21*(4), 1031–1033.

Vandesompele, J., K.˜De˜Preter, F.˜Pattyn, B.˜Poppe, N.˜Van˜Roy, A.˜De˜Paepe, and F.˜Speleman (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology*.

The GenABEL suite for genome-wide association analyses

Yurii S Aulchenko[1,2,3,⋆], on behalf of the GenABEL project

1. Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia
2. Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK
3. "Yurii Aulchenko" consulting, the Netherlands
⋆Contact author: yurii.aulchenko@gmail.com

Keywords: genome-wide association scans, SNP, human complex traits, statistical genomics

Genome-wide association (GWA) analysis is a widely recognized technique for identification of genomic regions (loci) in which changes in DNA sequence lead to changes in complex phenotype. In GWA scans, genomes of thousands of individuals are assessed by use of single nucleotide polymorphisms (SNP) arrays or whole-genome resequencing to gather information on hundreds of thousands to millions of genetic variants. The trait values of genotyped individuals are then tested for association with this genetic variation. During the last eight years, hundreds of loci for dozens of human common diseases and other complex traits were identified using GWA scans.

The **GenABEL** project aims to provide a free framework for collaborative, robust, transparent, open-source based development of statistical genomics methodology. In the framework of this project we have developed a suite of packages facilitating different semi-independent types of GWA analyses. The suite currently includes nine packages, of which seven are *R* libraries, such as the **GenABEL** for generic GWA quality control and analyzes, **MetABEL** for meta-analysis, **DatABEL** for large data sets management, **VariABEL** for identification of potentially interacting variants, and others. The packages are distributed under GPL or LGPL and are available at the **GenABEL** project home page, `http://www.genabel.org`.

Here, I will describe the **GenABEL** project in general and will also introduce some of the packages of the suite.

Making enzymes with R

TA Poulsen[1]

1. Novozymes A/S
*Contact author: tapo@novozymes.com

Keywords: HTS, ELISA, kinetics, data management

Novozymes is the world leader in industrial enzymes with products in detergency, food, feed, bio-fuel and more.

To make new and improved enzymes for these applications, we screen thousands of enzyme-variants per week. *R* is used liberally in the flow to transform, manage, analyze and understand the results.

The talk will give examples of high-throughput data processing and discuss some of the challenges in high-throughput data, including

- Experiment meta-data
- High-throughput data processing: server-side or client-side
- Local and global data storage
- Managing and sharing R-code
- Proprietary automation software
- Failure modes of high-throughput screening data

Use of molecular markers to estimate genomic relationships and marker effects: computation strategies in R

Filippo Biscarini[1*], Andrea Pedretti[1], Ulrike Ober[2], Malena Erbe[2], Hossein Jorjani[3], Ezequiel Nicolazzi[1] and Matteo Picciolini[1]

1. PTP (Parco Tecnologico Padano), Via Einstein - Loc. Cascina Codazza, 26900 Lodi (Italy)
2. Department of Animal Sciences, Georg-August-Universität, Albrecht-Thaer Weg 3, 37075 Göttingen (Germany)
3. Interbull Centre, Box 7023, S-75007 Uppsala (Sweden)

[*]Contact author: filippo.biscarini@gmail.com

Keywords: SNPs, matrix of genomic relationships, matrix inversion, parallel R, GPU

The *R* programming environment for data analysis has acquired popularity among scientists of different fields. Main reasons are the availability of many packages for statistical analyisis, the flexible programming syntax suited for tabular data, and the open source phylosophy. One limitation of *R* may lie in the speed of computation and in the capacity of dealing with large datasets. Large amounts of data are common in many areas of science and technology. Examples are high-density marker panels and whole-genome sequences (e.g. 30-odd million SNPs in the sheep genome) in bioinformatics and genomics. Computational strategies can be devised to deal with big data, such as parallel computing (partitioning a larger computation problem into sub-problems on different processors), and the use of the high-performing graphical processing unit (GPU, designed to handle the complex operations and computation load of the processing and rendering of images). Both approaches are especially suited to deal with issues of computational speed.

We explored the use of parallel computing and of the GPU processor to analyse large datasets in *R*. As working example, we chose an application in genomics: with a simulated set of 1000 genetic markers (SNPs, single nucleotide polymorphisms) and 4000 individuals Jorjani (2009), we set up the matrix $\boldsymbol{G}$ of genomic relationships between individuals VanRaden (2008). From this matrix we derived the SNP marker effects for a phenotype with mean 0 and standard deviation 10: this involves the calculation of the inverse of $\boldsymbol{G}$, a computationally intensive operation whose processing time is known to increase quadratically with the number of individuals. Initial calculations started with 1000 individuals: the population was then increased in steps of 100 individuals until reaching 4000 individuals. We compared the relative performances, in terms of computation speed, of the standard serial use of *R* with parallel and GPU computing. All calculations were performed on the same platform (machine and settings). The chosen problem involved the manipulation, inversion and multiplication of matrices and vectors: some of these operations could be vectorised, thus allowing for parallel computation. Parallelization was implemented in *R* using the package **parallel**. Cuda architecture and Cublas libraries were used for calculations on the GPU. Functions for the binding on the GPU and for some matrix operations (e.g. Cholesky factorization and Gauss-Jordan elimination) were written in *C* and wrapped in *R*. First results showed that, compared with the standard *R* computations, the parallel R and GPU implementations were faster on average by 17.4%, and 40.9%, respectively. The difference in computation speed was almost negligible with small matrices, but increased progressively with matrix size: for the largest matrices, the parallel R and GPU implementations were both $\sim 75\%$ faster than standard R. Besides, for more than 3200 individuals the computations in standard *R* were no longer possible, whereas the parallel R and GPU implementations could both reach the full data-size (4000 individuals). On average, the GPU implementation was faster than parallel R computation on the CPU by 25.8%.

References

Jorjani, J. (2009). A general genomics simulation program. *Interbull Bulletin* (40), 202.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science 91*(11), 4414–4423.

High Content Screening Analysis in R

Insa Winzenborg*, Pierre Ilouga

Discovery Informatics and Statistics, Evotec, Hamburg
*Contact author: insa.winzenborg@evotec.com

Keywords: High Content Screening, Multivariate Analysis, Linear Discriminant Analysis

High content screening (HCS) is a drug discovery method which is used to identify chemical substances (e.g. small molecules) that change the cell phenotype in a desired manner for a certain disease indication. Automated microscopy systems allow running these high content assays in a high throughput (up to 5,000 - 10,000 substances per day). The images acquired by automated microscopy are analyzed using image processing software, which extracts biologically meaningful parameters like the number of cells, cell roundness, cell area or fluorescence intensity. Many dozens of parameters could be obtained as a result e.g. on cell morphology, intracellular structures, and activation of intracellular signaling cascades.

With growing complexity of assays, it often happens that none of the extracted parameters is capable of reliably discriminating between biologically inactive and active substances. In this case, it is essential to combine several parameters into a robust multi-parameter readout that is straightforward to compute and at the same time easy to interpret. The linear discriminant analysis (LDA) is used to address this objective. Linear combinations of selected parameters are derived and lead usually to much better separations of the active and inactive control groups than the best parameter alone. In order to derive the multi-parameter readout, test plates are analyzed that only contain known inactive and active substances, i.e. negative and positive controls. They are randomly divided into a training and a test group. Linear combinations of different numbers of parameters are calculated on the training group (via exhaustive search or forward selection) and finally evaluated based on the test group. This process is repeated several times (internal cross validation). The quality of the obtained dimensionless multi-parameter readout is assessed by means of the so-called (multivariate) Z' factor (Kuemmel et al. 2010), which is a quality measure that incorporates the absolute difference of means and variability of negative and positive control groups.

However, there may be challenges in practical applications of this method. As an HCS campaign usually runs over several weeks it is desirable, at least for consistency and interpretability purposes, to apply the derived linear combination to all data generated throughout the campaign, even if uncontrollable factors lead to some measurement variation over time. For this reason, it is crucial to derive a robust readout combination, i.e. a combination that results in good Z' factors and therefore in a good separation of inactive and active substances in independent validations on several plates in several days (external cross validation).

The data analysis procedure, i.e. how many and which parameters are needed to obtain the best separation and what improvement of the Z' factor is achieved, as well as the graphical representation that shows the importance of parameters in the respective combinations was implemented in *R*.

The talk will present this HCS analysis method and its application to a real screening scenario.

References

Kuemmel, A., Gubler, H., Gehin, P., Beibel, M., Gabriel, D. & Parker, C. (2010). Integration of Multiple Readouts into the Z' factor for Assay Quality Assessment. *J Biomol Screen*, 1, 95–101.

rClr package – low level access to .NET code from R

Jean-Michel Perraud[1]

1. Commonwealth Scientific and Industrial Research Organisation, Australia
*Contact author: jean-michel.perraud@csiro.au

Keywords: interoperability, .NET, Mono, Common Language Runtime

rClr (http://r2clr.codeplex.com) is a package for *R* to access arbitrary *.NET* code executing on a Common Language Runtime implementation. The package can access the two main CLR implementations. On Windows® Microsoft's implementation is supported, and on this and other operating systems the cross platform implementation Mono is supported, although as of writing only Linux has been tested. **rClr** is the analogue for *.NET* to **rJava** (Urbanek, 2009) for the *Java* runtime. **rClr** complements and in part re-uses the already existing *R.NET* library (Abe, 2013) that makes *R* programmatically accessible to *.NET* programmers. The development of **rClr** is a personal endeavour motivated by work-related needs with complementary use of *R* and *.NET* code. Aside from ad-hoc file formats for data exchange there are existing programmatic solutions such as **rcom** but the underlying COM technology is effectively limited to Windows. Web Service based approaches are also possible and more platform agnostic. Both usually require additions or modification to existing .NET code to enable access from R, and the latter has a prohibitive performance penalty in some scenarios. **rClr** is designed to let R users access arbitrary .NET code (*C#*, *F#*, *VB.NET* and any other language that targets the CLR) without inherent need for addition or modification to this code. **rClr** has been used by the author to interactively test the correctness of a continental-scale, gridded spatial-temporal hydrological data assimilation method ported from *R* to *C#* to improve the runtime, scalability and integration with other systems. The seamless bi-directional conversion of the most common simpler *R* data types such as vectors is complete. Short to medium term work will center on the distribution via CRAN, runtime performance optimizations and the interoperability of more complex data types with no obvious or single equivalent in *.NET* such as data frames and S4 classes.

References

Abe, K. (2013). R.NET, http://rdotnet.codeplex.com

De Icaza, M. and others (2013). Mono project, http://mono-project.com

Urbanek, S. (2009). How to talk to strangers: ways to leverage connectivity between R, Java and Objective C, *Computational statistics* 24:303–311 DOI 10.1007/s00180-008-0132-x

Reproducible Research in Ecology with R: distribution of threatened mammals in Equatorial Guinea.

María V. Jiménez-Franco[1]*, Chele Martínez-Martí[2], José F. Calvo[1], José A. Palazón[1]

1. Departamento de Ecología e Hidrología, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain.
2. Wildlife Conservation International, New York Zoological Society, BronxPark, New York 10460, USA.
*Contact author: mvjimenez@um.es

Keywords: R raster package, reproducible research, markdown, occupancy models, R knitr package

Studies using data from different sources (e.g., field data and Geografical System Information, GIS) and different working levels or steps (e.g., obtaining cartographic variables, statistical analysis and maps species occurrence), may have some problems to perform the methodological framework easily. This context was arises in a conservational study that aimed to map species-specific occurrence probability identifying priority areas for conservation of seven threatened species in Equatorial Guinea (Martínez-Martí, 2011). The application of different packages and their combination through *R* (R Core Team, 2013) allow us to standardize repetitive processes and compile important scripts for a better understanding of all members of the working team and for future applications. Here, we use Reproducible Research to combine statistical analyses and spatial data, using **raster** (Hijmans and Van Etten, 2012) and **kntir** packages (Xie, 2013) and *markdown* language (http://daringfireball.net/projects/markdown/).

We obtained a useful document, which uses Reproducible Research to combine the spatial data obtained through GIS and the analytical processes needed to obtain the main areas of species occurrence in Equatorial Guinea. This document can be read and understood easily after a long period of time by the same authors and other researchers, which indicates that it can be reused or modified for other similar studies. Although the first time preparing the document with *markdown* language may take some time, we suggest that this working method is a useful tool that facilitates the learning and the work in the *R* proceedings. This process of compiling the methods in a document could be applied not only for ecologists and researchers of other scientific areas but also for students in their degrees and masters.

References

Hijmans, R.J., Van Etten, J. (2012). Geographic analysis and modeling with raster data. URL http://cran.r-project.org/web/packages/raster/raster.pdf.

Martínez-Martí, C. (2011). The leopard (*Panthera pardus*) and the golden cat (*Caracal aurata*) in Equatorial Guinea: A national assessment of status, distribution and threat. Annual report submitted to Panthera/Conservation International.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Xie, Y. (2013). knitr: A general-purpose package for dynamic report generation in R . R package version 1.1, URL http://yihui.name/knitr

Using R for Mapping the Spatial Extent of Meteorological and Hydrological Drought Events

Jiří Kadlec[1,2,*] , Pavel Treml[2,3]

1. Environmental Informatics Research Group, Aalto University, Finland
2. Charles University Prague, Department of Physical Geography and Geoecology, Czech Republic
3. TG Masaryk Water Research Institute, Prague, Czech Republic
*Contact author: jiri.kadlec@aaalto.fi

Keywords: Drought, Map, Interpolation, Spatial Analysis

Drought is defined as an extended period of deficient water supply in a region. Typically the first phase is meteorological drought (deficit precipitation) that is followed by long-term hydrological drought. The drought event develops in space and time with an alert, emergency and rehabilitation phase. The spatial extent of a drought event shown on a map depends on the selected drought indicator, density of available meteorological and hydrological observation sensors, interpolation method and time series smoothing. For a drought expert it is important to view the development of a drought event as a series of maps or a map animation that for each day overlays station locations, region boundaries, major rivers, and the calculated drought severity grid. For comparing maps it is required that all grids share the same color ramp and color break intervals. For automating the creation of drought we present a solution approach in *R* that uses the **sp**, **maptools** and **gstat** packages. First, the drought expert selects a function that transforms a time-series of meteorological or hydrological measurements (air temperature, precipitation or discharge) to a time-series of drought index values. Existing *R* packages such as the standardized precipitation index (**spi**) may be re-used for developing this function. The meteorological measurements may be obtained from user-defined text files or from online data sources including WaterML and Sensor Observation Service. Second, for each station in the area of interest, an automatic script reads the measurements time series and calculates the drought index. Third, for each day in the time period of interest, the value of the drought index at all station is read and a grid is interpolated using a pre-defined geostatistical or deterministic interpolation method such as kriging or inverse distance weighted. Finally a map layout with country boundaries, rivers, station locations, labels and color key is added to the grid and each map is saved to an image. This allows easy creation of animations from the saved images. The advantage of the *R* approach compared to a desktop geographic information system (GIS) solution is that *R* simplified the automation of the map creation workflow that enables drought researchers to quickly explore the effects of selected drought indicator parameters, time smoothing and interpolation methods on the estimated spatial extent of the developing drought event.

Three-component decomposition of coal spectrum in *R*

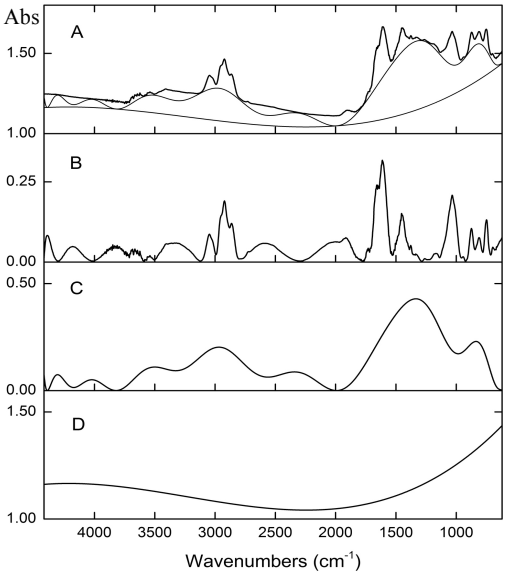
Yuri Possokhov[1*]

1. Eastern Research & Development Institute of Coal Chemistry (VUKHIN), Yekaterinburg, Russia
*Contact author: possokhoff@gmail.com

Keywords: Spectroscopy of coals, Spectrum decomposition, Polynomial background, Chemometrics

R with its *CRAN* is an excellent tool for building automated high-accuracy tests of spectroscopic data in Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectrometry of coals. *R*'s power is revealed by mix of truly ease in deployment of complex algorithms and its code clarity. That is why, *R* is chosen as a language and development environment for building *Spectrotest-SDK* – the special development kit, which provides functions and S3-classes for rapid automation of DRIFT-data processing tasks.

The identification and further interpretation of broad bands in DRIFT-spectra of coals has a long-term story, which arises from various hypotheses of coal structure. Given these circumstances we searched for simple, yet effective approach to divide spectral data into strictly localized and mainly delocalized parts. We found the non-quadratic cost function technique [1] to be the starting point to fit a polynomial as the delocalized part, i.e. the background. As a result, we authored *R*-function `fons` included in *Spectrotest-SDK* to programmatically realize the "coal-orientied" modification of the proposed technique. Herewith, the *R*-code of the authored function looks rather plain and simple due to function `pseudoinverse` located in the remarkable package **copcor**.



Three-component decomposition of DRIFT-spectrum of bituminous coal: A – polynomial fit; B – monomer; C – polymer; D – methodical background

Thoroughly looking through numerous combinations of tuning parameters for the function `fons`, we found those values that assisted in three-component decomposition of DRIFT-spectra of more than 100 coals at different metamorphic stages. Those components may be consecutively treated as spectra of monomer (localized) and polymer (delocalized) parts of chemical structure and methodical background.

So, we will describe the latest practical experience of implementing this decomposition in the light of application of chemometric tools for coal calibration in *R* [2].

References

[1] Vincent Mazet et al. (2005). Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems 76*, 121–133.

[2] Possokhov Yu.M., Popov V.K., and Butakova V.I. (2010). Application of chemometric tools for coal calibration by DRIFT spectroscopy. In *Modern Methods of Data Analysis 2010, The Seventh Winter Symposium on Chemometrics, (Saint Petersburg, Russia)*, pp. 50–53.

Method of comparison of actions of the liquidators of the accident on Chernobyl Nuclear Power Plant on the basis of fragmentation of their routes and encryption it in a form similar to the DNA

Konstantin Chizhov[1,*] , Ilya Kudrin[1], Elena Bakhanova[2], Petr Bondarenko[2], Ivan Golovanov[1], Vladimir Drozdovitch[3],Vladimir Chumak[2], Viktor Kryuchkov[1]

1. Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, RF Ministry of Health and Social Development. 46, Zhivopisnaya St., Moscow, 123182, Russian Federation
2. Radiation Protection Institute ATS Ukraine, Kiev
3. DHHS, NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Boulevard, Bethesda, MD 20892, USA
*Contact author: nicemind@ya.ru

Keywords: Radiation protection, dose comparison, DNA, sets analysis, Venn-Euler diagrams.

Using language *R* we have developed algorithms for comparison of profiles of the liquidators of the accident on Chernobyl Nuclear Power Plant (ChNPP). Each profile is a questionnaire, that consists of a set of human actions that are recorded in a special format - for every action is defined date, time, place and the protective factors that weaken the dose. Also, using the method RADRUE[1], restored radiation environment for each region and each day.

An example of simple element (in our terminology - frame) from the questionnaire is movement from Pripyat to the industrial site of ChNPP, or working in one of the rooms. One questionnaire can include up to several thousands of such actions.

It is necessary to fragment frames to the extent that they represent the basic objects in terms of calculating the dose, i.e. divide them up to elementary level, when the dose rate is constant for all fragment. Then, the dose that corresponds to a fragment of the frame is equal to the product of the dose rate and the time spent on this fragment. Thus we have a chain of actions of the liquidator, presented as a consistent set of codes similar to the DNA chain. And then, using the techniques of genetic analysis of package **GenoPlotR**[2], we get a graphical representation of the coincidence of the two profiles.

Also we made a comparison of questionnaires as a comparison of Sets[3] and quantify their intersection. We have calculated Measure of distinction between two questionnaires and Soft Measure of distinction, that not takes into account the location of the liquidator. Localization of liquidator has absolutely no importance in terms of epidemiology, as the dose to be received by him for a certain time in the selected room in any case will be the same. The comparison results are presented in the form of Venn-Euler diagrams[4].

As a result of the work we have checked profiles of 28 emergency workers who are filled questionnaire immediately after accident and 20 years later. Based on the comparison of methods listed above we received in graphic and text form the common and various elements of their questionnaires. Through this analysis, we can find the exact place where the liquidator reported incorrect information or to quantify the level of forgetfulness. Also, this method can be used to

Differential expression analysis of RNA-seq data at base-pair resolution in multiple biological replicates

Alyssa Frazee[1], Leonardo Collado-Torres[1], Andrew Jaffe[1,3], Sarven Sabunciyan[2], Jeffrey T. Leek[1,⋆]

1. Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA
2. Stanley Division of Developmental Neurovilogy, The Johns Hopkins School of Medicine, Baltimore, MD, USA
3. Lieber Institute for Brain Development, Maltz Research Laboratories, Baltimore, MD, USA

⋆Contact author: jtleek@gmail.com

Keywords: RNA-seq, derfinder, differential expression

Since the invention of microarrays, measuring genome-wide gene expression has become a common experiment performed by molecular biologists and clinicians. Detecting differentially expressed genes is arguably the most common application of this technology. RNA-sequencing (RNA-seq) is a more flexible technology for measuring genome-wide expression that is rapidly replacing microarrays as costs become comparable. RNA-seq experiments produce billions of short sequences, obtained from individual RNA transcripts, referred to as reads. Current statistical methods for differential expression analysis based on RNA-seq data fall into two broad classes based on how they summarize the information in the reads: (1) methods that count the number of reads within the boundaries of genes previously published in databases and (2) methods that attempt to reconstruct full length RNA transcripts. Both methods have limitations. The first cannot discover differential expression outside of previously known genes, which negates one of the novel aspects of the technology. While the second approach does possess discovery capabilities, the existing implementation grossly underestimates the uncertainty introduced during the summary step and thus cannot reliably detect differential expression. Frazee et al. (2013) proposed a statistical pipeline that preserves the discovery capability of the second approach while achieving similar stability to the first approach. It achieves this by measuring the number of reads overlapping each individual base-pair, then grouping consecutive base-pairs with common differential expression patterns into differentially expressed regions (DERs). Novel regions and regions that overlap known genes are then labeled for downstream use. This approach is referred to as DER Finder, is implemented in *R* as the **derfinder** package and is available on GitHub.

We compare the DER Finder approach to leading competitors with a large data set with multiple biological replicates. Large data sets like this one pose new challenges when implementing the methods. We thus optimized pre-processing steps for the DER finder pipeline to reduce computational requirements and the overall analysis time.

References

Frazee, A. (2013). derfinder. https://github.com/alyssafrazee/derfinder.

Frazee, A., S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2013). Differential expression analysis of rna-seq data at base-pair resolution. *Manuscript in revision*.

Statistical inference for Hardy-Weinberg equilibrium with missing data

Jan Graffelman[1,⋆]

1. Department of Statistics and Operations Research, Polytechnic University of Catalonia
⋆Contact author: jan.graffelman@upc.edu

Keywords: Single nucleotide polymorphism, inbreeding coefficient, multiple imputation.

In genetic association studies polymorphisms are usually tested for Hardy-Weinberg equilibrium (HWE), typically by using chi-square or exact tests. Significant deviation from Hardy-Weinberg equilibrium can be indicative of genotyping error or marker-disease association. Genetic markers often have a considerable amount of missing values. Genotyping platforms apply clustering/classification algorithms to allele intensities in order to classify individuals as AA, AB or BB. Missing values arise if the algorithm is unable to assign a genotype to an individual for the given allele intensities. When markers are tested for equilibrium, the missing values are usually discarded. This can lead to biased inference about HWE when genotype data is not missing completely at random. In this contribution we propose to impute missing genotypes using a multinomial logit model. In order to impute the missings, the model uses allele intensities and information from neighbouring markers. We perform multiple imputation and estimate the inbreeding coefficient for each imputed dataset. Next, we use Rubin's pooling rules (Little and Rubin, 2002) to combine inbreeding coefficients over all imputed datasets. The result is a test for HWE that can take missing data into account.

We applied our test to an empirical database of single nucleotide polymorphisms possible related to colon cancer. Missing genotype data turned out not to be missing completely at random. Markers with significant deviations from Hardy-Weinberg equilibrium typically showed a lack of heterozygotes. When missing data were imputed with the multinomial logit model, missings were often imputed as heterozygotes. Accounting for missing qualitatively changed up to 17% of all test results for HWE.

All computations were performed in *R*. We used graphical and inferential tools (Graffelman and Morales-Camarena, 2008) from the package **HardyWeinberg** and facilities for the imputation of missings from the **mice** package (Buuren and Groothuis-Oudshoorn, 2011). Special functions for testing for HWE in the presence of missings were developed and added to the **HardyWeinberg** package (Graffelman, 2012).

References

Buuren, S. v. and K. Groothuis-Oudshoorn (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software 45*(3), 1–67.

Graffelman, J. (2012). *HardyWeinberg: Graphical tests for Hardy-Weinberg equilibrium*. R package version 1.5.1.

Graffelman, J. and J. Morales-Camarena (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity 65*(2), 77–84. doi: 10.1159/000108939.

Little, R. J. A. and D. B. Rubin (2002). *Atatistical analysis with missing data* (second ed.). New York: John Wiley & sons.

What did we learn from the IMPROVER Diagnostic Signature Challenge?

Adi L. Tarca[1,2,*], Roberto Romero[2], Julia Hoeng[3], Manuel Peitsch[3], Gustavo Stolovitzky[4]

1. Department of Computer Science, Wayne State University, Detroit, MI, USA
2. Perinatology Research Branch, NICHD/NIH, Bethesda, MD, and Detroit, MI, USA
3. Philip Morris International, Research & Development, Neuchâtel, Switzerland
4. IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

*Contact author: atarca@med.wayne.edu

Keywords: outcome prediction, microarray data, crowdsourcing

The IMPROVER Diagnostic Signature Challenge (DSC) [1,2], was designed by scientists from Philip Morris International's (PMI) Research and Development department and the IBM's Thomas J. Watson Research Center, with the goal of assessing the robustness of methods currently in use for outcome prediction using high-dimensional biological data. In this double blind crowdsourcing competition funded by PMI, public microarray datasets were suggested for developing prediction models in four disease areas (5 endpoints total). The scientific community participated in the challenge, with 54 teams submitting predictions on new sets of samples generated by organizers. The predictions were ranked using three different performance metrics. The team AT & RR received the best performing entrant award, being ranked 2nd in three of the four scored sub-challenges and 12th on the fourth one.

In this work, we will present the main results from the IMPROVER DSC including ranking stability analyses of the participating teams and identification of modeling factors that explained the models success and performance variability. The approach of the best overall team is also presented including an *R* package called **maPredictDSC**, that implements their classification pipeline. The main function of the package starts with raw microarray data files and a class label for each training sample, and returns fitted models and their predictions on the test samples. In addition, the package allows to explore 26 other combinations of preprocessing, features selection and classification methods. Using performance data from the 27 different models produced by **maPredictDSC** as well from the models submitted in the challenge we have concluded among others that: i) no fixed classification pipeline works best for all datasets ii) the endpoint explains most of the variability in the performance data, of iii) the importance of various steps involved in the classification is dataset and metric dependent iv) classical discriminant analysis methods seemed to perform at least as well as emerging prediction algorithms specifically designed for high-dimensional data provided that proper tuning was made to the specificities of each dataset.

The use of crowdsourcing to validate research building blocks of interest for both academia and the industry proved to be promising by allowing a large body of computational work to be conducted in a few months rather than years.

References

1. Meyer P et al. (2011). Verification of systems biology research in the age of collaborative competition. *Nat.Biotechnol. 29*(*9*), 811-815.
2. Meyer P et al. (2012) Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. *Bioinformatics. 28*(*9*), 1193-1201.
3. Tarca, A. L, Than, N. G., Romero, R. (2013) Methodological Approach from the Best Overall Team in the IMPROVER Diagnostic Signature Challenge, *Systems Biomedicine* submitted.

Deciphering the tRNA operational code - using R

Tal Galili[1], Shaul Shaul[2], Yoav Benjamini[1]

1. Department of Statistics and Operations Research, Tel Aviv University, Israel.
2. Department of Zoology, Tel Aviv University, Israel.
***Contact author: Tal.Galili@gmail.com**

Keywords: tRNA, amino acids, operational RNA code, rpart, dendrogram

The talk will deal with the unique properties of the RNA code that governs the charging of the transfer-RNA (tRNA) molecule so that it will bring the appropriate amino-acid to the Ribosome.

The analysis was performed on 3936 tRNA sequences from 86 Archaea species using R. We employed various existing facilities for performing data importing (**Biostrings, XML, RCurl**), cleaning and preparation (**plyr**, **reshape**), classification and regression trees and cross-validation (**rpart**), clustering (**cluster**), visualization (**Graphics**, **lattice**, **seqLogo, colorspace**), reproducible research (**Sweave**, **knitr**, **xtable**, **Hmisc, installr**) etc.
In addition, we self-developed (or implemented) algorithms for manipulating dendrograms objects for tasks such as the comparing of hierarchical clusters and new plotting options (all of which are intended to be released in the near future in the **dendextend** package).

This talk is much-updated follow-up for the 2010 useR talk I gave on the same topic (see ref 1). In this talk I will provide the brief biological background that is needed in order to understand the relevant questions and discoveries, and present how we used R's various packages and statistical methods, while devising and implementing new methods, in order to support our investigation.

References

1. Tal Galili, Shaul Shaul, Yoav Benjamini (2010), "Analyzing the Operational RNA Code for Amino Acids - Using R", a talk from useR2010 - http://user2010.org/abstracts/Galili+Shaul+Benjamini.pdf
2. Shaul S, Berel D, Benjamini Y, Graur D.(2010), "Revisiting the operational RNA code for amino acids: Ensemble attributes and their implications", RNA. 2010 Jan;16(1):141-53. Epub 2009 Dec 1.

[1] Department of Statistics and Operations Research, Tel Aviv University, Israel.
[2] Department of Zoology, Tel Aviv University, Israel.

Big Data and Reproducibility – Building the Bridge

Andreas Leha

University Medical Center Göttingen – Department of Medical Statistics – Core Facility Statistical Bioinformatics
Contact: andreas.leha@med.uni-goettingen.de

Keywords: Reproducible Research, Big Data, Version Control

This is the "decade of big data" [2]. In the field of sequence analysis, for example, datasets the size of several gigabytes are frequently encountered. The availability of big datasets, though, brings about new challenges not only for the analysis methods but also for the handling of the data – especially in connection reproducible research [4, 3].

Since reproducibility demands the data to be present in a specified version, version control systems for both code and data are of great help to the researcher. Until recently, the available version control systems failed to handle big data well. As a result, massive datasets are seldom put under version control and, thus, are often kept separate from the project.

The novel *git-annex* [1] extension to the *git* version control system closes this gap. As the original target of *git-annex* were multimedia files, this system is tailored with a special focus on big data files.

We present a wrapper from *R* to *git-annex* that makes it easy to handle big datasets from within the analysis project. With this method, also big data finally becomes part of the project repository and is easily accessible. In addition, the content of the large files can be safely moved away from the working directory and backups get automatized.

References

[1] Hess, J. (2012). git annex. http://git-annex.branchable.com/.

[2] Ranganathan, S., C. Schönbach, J. Kelso, B. Rost, S. Nathan, and T. W. Tan (2011, Nov). Towards big data science in the decade ahead from ten years of incob and the 1st iscb-asia joint conference. *BMC Bioinformatics 12 Suppl 13*, S1.

[3] Schulte, E., D. Davison, T. Dye, and C. Dominik (2012, 1). A multi-language computing environment for literate programming and reproducible research. *Journal of Statistical Software 46*(3), 1–24.

[4] Xie, Y. (2012). *knitr: A general-purpose package for dynamic report generation in R*. R package version 0.3.

Topology-based Hypothesis Generation on Causal Biological Networks using igraph

Robert Ness[1*], **Halima Bensmail**[3], **Olga Vitek**[1,2]

1. Department of Statistics, Purdue University, West Lafayette, Indiana, USA
2. Department of Computer Science, Purdue University, West Lafayette, Indiana, USA
3. Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar

[*]Contact author: nessr@purdue.edu

Keywords: interaction networks, causal graph, proteomics, igraph

A causal biological network as a type of molecular interaction network constructed from biomedical literature. The nodes are molecular events, such as a change in abundance of a protein. Directed edges between two events imply a causal relationship between them. The edges have a directional attribute in terms of increase or decrease. For example, catalytic activity of enzyme A increases the abundance of protein B. Each edge is annotated with a reference to a publication from the biomedical literature that shows experimental validation for the edge. A key difference from similar network models such as Bayesian networks is that the directional edge attribute does not have a numerical magnitude (eg. a probability). This is because the edges each come from different experimental contexts, so it is difficult to establish a basis for numerical comparison.

Causal networks are used to generate and prioritize mechanistic hypotheses for what is observed in experimental data. Specifically, we look at genomics and proteomics data with treatments and a control. Statistics for each feature (eg. gene probe measurement or protein peptide spectra) in the data are mapped to nodes in the network. Typically, the statistic is a 1 if the features measurements changed significantly across treatments, and 0 otherwise. Other nodes in the network are evaluated as mechanistic hypotheses for what was observed in the data, based on the directed paths in the network. High ranked hypotheses can then be validated experimentally. This approach is particularly useful in drug discovery and drug repositioning, for example, because of the potential to identify molecular mechanisms susceptible to drug intervention.

The question we ask is what is the best path-based scheme for ranking nodes in the network. Common practice is to use the shortest-paths algorithm. A node is ranked highly if there is a large number of shortest paths from itself to the nodes with a 1. The problem with this approach is that fails to incorporate broader network topology, in terms of the total number of all directed paths between the source node and the target nodes, rather than just the number of shortest paths. We test an alternative approach that uses a Markov random walk-based algorithm similar to Pagerank. This approach counts the total number of paths between a source and target node using a weighting function that assigns higher weights to more direct paths. We compare the performance of this algorithm to shortest-paths and demonstrate its effectiveness for hypothesis generation in systems biology experiments.

We conduct our analysis using the R package **igraph**. **igraph** is an open source software package for creating and manipulating undirected and directed graphs with implementation as a library in R, as well as other programming languages. Though causal biological networks can be small (100s of nodes), we are only interested in situations when they are large (10,000+ nodes), such when knowledge in a knowledgebase is being used to analyze data. The size of the network significantly impacts the speed of most network analysis algorithms. **igraph** contains several functions implemented in C that allow for fast analysis of large networks, including shortest-paths and other network topology algorithms.

Funding: This work was supported by the Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar

Hansel: A Deducer Plug-In for Econometrics

R. Scott Hacker[1,*]

1. Jönköping International Business School, Jönköping, Sweden
*Contact author: Scott.Hacker@jibs.hj.se

Keywords: GUI, *R*, **Deducer**, econometrics, time series, panel data

This paper discusses the development of a **Deducer** plug-in, referred here to as **Hansel**, that can deal with techniques typically found in undergraduate courses in econometrics, along with some more advanced econometric techniques (the final package name should be something like **DeducerHansel**). Currently the **Deducer** package (Fellows, 2012) 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 **Hansel** 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 techniques. **Hansel** 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

Ian Fellows (2012). Deducer: A Data Analysis GUI for R. Journal of Statistical Software, 49(8), 1-15.
URL http://www.jstatsoft.org/v49/i08/.

Robust standard errors for panel data: a general framework

Giovanni Millo[1,⋆]

1. Generali Research and Development
⋆Contact author: mailto:giovanni_millo@generali.com

Keywords: Standard errors, Panel data, Heteroskedasticity, Clustering, Persistence

A comprehensive, modular and flexible framework is described for estimation of robust standard errors in panel data. Heteroskedasticity and autocorrelation robust estimators in the tradition of White (1980) and Arellano (1987) are brought together with the "SCC" mixing-fields based estimator of Driscoll and Kraay (1998), the unconditional "PCSE" estimator of Beck and Katz (1995), and the double-clustering approach of Cameron et al. (2011) and Thompson (2011), trying to bring together the applied literatures in macroeconometrics, finance, political science and accounting by demonstrating the common features of these apparently different approaches. The covariance estimators are integrated in package **plm**, complying with the standards of package **lmtest** and allow robust specification and restriction testing over a number of different panel models.

References

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics 49*(4), 431–434.

Beck, N. and J. N. Katz (1995). What to do (and not to do) with time-series cross-section data. *American political science review*, 634–647.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*(2), 238–249.

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics 80*(4), 549–560.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics 99*(1), 1–10.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Rsiopred: An R package for forecasting by exponential smoothing with model selection by a fuzzy multicriteria approach

C. Bergmeir[1,*], J. M. Benítez[1], J. Bermúdez[2], J.V. Segura[3], E.Vercher[2]

1. DECSAI, DiCITS Lab, SCI2S group, CITIC-UGR, Universidad de Granada, Granada, Spain
2. Dpto. Estadística e Investigación Operativa. Universitat de València.
3. Centro de Investigación Operativa. Universidad Miguel Hernández.
*Contact author: c.bergmeir@decsai.ugr.es

Keywords: Forecasting, Time series, Exponential smoothing, Non-linear programming

Exponential smoothing is one of the oldest, most widely used and most successful forecasting procedures (Goodwin 2010). Its potential applicability has recently been even increased with the introduction of a complete modeling framework incorporating innovations state space models, likelihood calculation, prediction intervals and procedures for model selection (Ord *et al.* 1997; Bermúdez *et al.* 2007; Hyndman *et al.* 2008; Vercher *et al.* 2012). In *R*, some basic models of exponential smoothing are available through the `HoltWinters()` function in the **stats** package, and the function `ets()` from the **forecast** package provides a larger set of models and optimization possibilities (see also the CRAN Task View for Time Series Analysis).

We present the package **Rsiopred**, which implements the SIOPRED forecasting procedure (Bermúdez *et al.* 2006, 2008). The procedure is based on a model that works with three components: mean level, damped trend and seasonality factors. The seasonal effects can be modeled using either additive or multiplicative forms. Our methodology unifies the phases of estimation and model selection just into an optimization framework which permits the identification of robust forecasts. Although the incorporation of the initial values as decision variables of the optimization problem increases the dimension of the related non-linear programming problems, the use of suitable optimization tools in the estimation analysis has been very fruitful, providing accurate forecasts.

A very important part of the method is the non-linear solver used. In its original version, SIOPRED uses a proprietary solver (Frontline Systems Inc.). **Rsiopred** implements, besides the possibility to use this proprietary solver, interfaces to available solvers in *R*, namely **Rsolnp**, **ipoptr**, and **Rdonlp2**. Various performance issues are tackled by the use of *C++* code together with **Rcpp**. We analyze performance of the different solvers and the ex-post forecasts for some real time series datasets in order to illustrate our approach.

References

Bermúdez, JD,; Segura, JV.; Vercher, E. (2006) A decision support system methodology for forecasting of time series based on soft computing. *Computational Statistics & Data Analysis* 51, 177-191.

Bermúdez, JD.; Segura, JV.; Vercher, E. (2007) Holt-Winters forecasting: an alternative formulation applied to UK Air passenger data. *Journal of Applied Statistics* 34, 1075-1090.

Bermúdez, JD.; Segura, JV.; Vercher, E. (2008) SIOPRED: a prediction and optimization integrated system for demand. *TOP* 16, 258-271.

Ord, JK.; Koehler, AB.; Snyder, RD. (1997) Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* 92, 1621-1629.

Goodwin, P (2010). The Holt-Winters approach to exponential smoothing: 50 years old and going strong. Foresight: The International Journal of Applied Forecasting, 19:30–33.

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) *Forecasting with exponential smoothing. The state space approach*. Springer, Berlin.

Vercher, E.; Corberán-Vallet, A.; Segura, JV.; Bermúdez, JD. (2012) Initial conditions estimation for improving forecast accuracy in exponential smoothing. *TOP* 20, 517 - 533.

AutoSEARCH: Automated General-to-Specific Model Selection

Genaro Sucarrat[1,⋆]

1. BI Norwegian Business Schoool
⋆http://www.sucarrat.net/, genaro.sucarrat@bi.no

Keywords: General-to-specific, model selection, AR-X, log-ARCH-X, volatility

General-to-Specific (GETS) modelling starts with a General Unrestricted Model (GUM) that is validated against a chosen set of chosen misspecification tests. Next, multi-path simplification is undertaken by means of backwards elimination, where each regressor-elimination is checked against the chosen misspecification tests, and by a BackTest (BaT) – also known as a parsimonious encompassing test – against the GUM. Simplification stops when there are no more insignificant regressors, or when the remaining possible deletions either do not pass the misspecification tests or the BaT. Since simplification is undertaken along multiple paths, this usually results in multiple terminal models. An information criterion is thus used to choose among them. **AutoSEARCH** undertakes automated GETS model selection of linear regression models – possibly including Autoregressive (AR) terms and exogenous (X) conditioning variables, of log-ARCH-X models, or in both.

The tuning parameter in GETS model selection is the significance level, which corresponds to the targeted irrelevance proportion. In other words, the proportion of irrelevant variables that are retained will – on average – equal the chosen significance level. So a low significance level should be chosen if a parsimonious model is desired, or if the starting model contains many regressors. Asymptotically, the final selected model contains the Data-Generating Process (DGP) with probability 1.

Most financial models are highly non-linear and require complex optimisation algorithms and inference strategies in empirical application. Indeed, this may even become an obstacle to automated financial GETS modelling, as for example argued by Granger and Timmermann (1999) regarding automated GETS modelling of financial volatility. A solution to this problem is provided in Sucarrat and Escribano (2012), where GETS model selection is studied in a log-ARCH-X model. The current version of **AutoSEARCH** is a refinement of the code developed for that project.

GETS model selection is closely related to – but not the same as – the GETS methodology, see Campos et al. (2005) for a comprehensive overview, and Mizon (1995) for a concise one.

References

Campos, J., D. F. Hendry, and N. R. Ericsson (Eds.) (2005). *General-to-Specific Modeling. Volumes 1 and 2*. Cheltenham: Edward Elgar Publishing.

Granger, C. W. and A. Timmermann (1999). Data mining with local model specification uncertainty: a discussion of Hoover and Perez. *Econometrics Journal 2*, 220–225.

Mizon, G. (1995). Progressive Modeling of Macroeconomic Time Series: The LSE Methodology. In K. D. Hoover (Ed.), *Macroeconometrics. Developments, Tensions and Prospects*, pp. 107–169. Kluwer Academic Publishers.

Sucarrat, G. and Á. Escribano (2012). Automated Model Selection in Finance: General-to-Specific Modelling of the Mean and Volatility Specifications. *Oxford Bulletin of Economics and Statistics 74*, 716–735.

Driving *R* to the air quality industry. NanoEnvi Analyst: a tool for designing large-scale air quality plans for improvement in ambient air quality

A. Alija[1⋆], M. N. Gonzalez[1], P. Fernandez Acebal[2] and J. Blanco[1]

1. Ingenieros Asesores S.A. Parque Tecnologico de Asturias. Parcela 39, 33428. Principado de Asturias. Spain
2. Dpto. Fsica. Facultad de Ciencias 33007. Universidad de Oviedo. Principado de Asturias. Spain

⋆Contact author: aab@ingenierosasesores-sa.es

Keywords: Air quality, Air pollution, **openair**, **rstudio-server**, **shiny**, data mining, **lubridate**.

The importance of monitoring and controlling air pollution in both outdoor and indoor environments is well established in both the literature and in legislation, especially for urban areas. Air pollutants have harmful effects on the environment and living organisms. In many countries, air pollutant concentrations are still above the legal and recommended limits that are set to protect the health of European citizens. The year 2013 will be the Year of Air and the European Commission is preparing a review of EU air legislation [Age11]. The EU DIRECTIVE 2008/50/EC on ambient air quality and cleaner air [Uni08] for Europe establish that air quality plans should be developed for zones and agglomerations within which concentrations of pollutants in ambient air exceed the relevant air quality target values or limit values.

In this work, we present an intensive use of *R* and more concretely package **openair** [CR11] as a tool for developing air quality plans. In particular the air quality plan presented here is focused on particulate matter called PM_{10} what is one of the most air quality concern in Europe now. Package **openair** lead out to analyse long historical time series of air pollution data coming from a large urban region somewhere [pri]. Long historical series of data have the inconvenience of they are very difficult to analyze for several reasons. Perhaps, some of these reasons are: first, air pollution data is a data extremely correlated in time and second, air pollution data is a complex data which require a very optimized visualization tools to carry out even the simplest analysis. This work presents a clear example of how R meets the demands and requirements of modern data-driven businesses. In this sense, Envira company shows here a real case study on the air quality industry where NanoEnvi Analyst (NEA) offers our customers an intelligent data-mining tool for air quality data interpretation. NEA facilitates decision-making and the development of effective strategies for air pollution mitigation.

References

[Age11] EEA European Environment Agency. *SOER2010*, 2011.

[CR11] David Carslaw and Karl Ropkins. *openair: Open-source tools for the analysis of air pollution data*, 2011. R package version 0.4-0.

[pri] private. To preserve the confindenciality of the source of the data we will keep the location anonimus.

[Uni08] European Union. Directive 2008/50/ec on ambient air quality and cleaner air for europe. Technical report, 2008.

Sequential Design of Experiments for model selection: an application to the energy sector

Daniele Amberti [1,2,*] , Michele Giordani[1], Alessandra Padriali[1]

1. i4C S.r.l., www.i4CAnalytics.com
2. Torino R net, www.TorinoR.net
*Contact author: daniele.amberti@i4canalytics.com

Keywords: design of experiments, model selection, additive models, time series, energy, large forecasting problems

The problem of choosing a good forecasting model, and then estimate its parameters, is well known among forecast practitioners. In the energy sector, a number of models like Linear Models, ARIMA, Generalized Additive Models and Neural Networks have been used to forecast gas consumption or electricity load. The energy sector is of particular interest because of its complex seasonality, the number of predictors involved in the process (e.g. lagged values, calendar effects, weather data and economic cycle) and the high number of time series that have to be forecasted. In this presentation we concentrate on the application of Generalized Additive Models to forecasts in the gas sector and we propose a procedure to search for best forecasting models for a set of similar time series.

We make use of **mgcv** for Generalized Additive Models and several packages and functions from the **ExperimentalDesign** Task View to construct a strategy for the Sequential DoE. Package **forecast** and work of Hyndman R. in the energy sector is relevant as a background.

Emission inventory supported by *R*: dependency between calorific value and carbon content for lignite

Damian Zasina[1,2*], Jarosław Zawadzki[1]

1. Warsaw University of Technology, Faculty of Environmental Engineering, 00-653 Warsaw, 20 Nowowiejska Str.
2. Insititute of Environmental Protection – National Research Institute, National Centre for Emission Management, 00-805 Warsaw, 132/134 Chmielna Str.
*Contact author: damian.zasina@gmail.com or damian.zasina@kobize.pl

Keywords: *R*, emission inventory, emission factor, carbon dioxide, GHG

The paper presents possible usability of *R* as a supporting tool for tasks, elaborated primarily for the purposes related to greenhouse gases (GHGs) emission inventory and management on the level of country. In the context of climate change matters and international obligations taken by Poland as a result of ratifying the United Nations Framework Convention on Climate Change (UNFCCC) there exists requirement of compiling and submission the annual national emission inventory of GHGs.

Opencast mining in Poland and combustion of lignite is still one of the most important ways of obtaining electrical power. According to statistical data, elaborated by The Energy Market Agency (EMA), almost 32% [EMA, 2013] of produced electricity is based on combustion of the lignite. The Polish case study showed that lignite will be significant source of energy till 2040 [Kasztelewicz, Kozioł, 2007].

For public power sector, as the most important source of the emission of carbon dioxide (CO_2), the estimation of emission of CO_2 is based on correlation between the calorific value [MJ/kg] and carbon content in fuel [%]. This dependency is elaborated as a linear function of the correlation, based on measurements [Fott, 1999; Stefanović et al., 2012a; Stefanović et al., 2012b]. Emission of CO_2 is the result of straightforward conversion of the carbon content into emission value.

The case study aims at estimation of the national mean carbon content and comparison of current (*status-quo*) analysis of dependency between calorific value and carbon content with the new one, developed using *R* as a basic statistical tool. Elaborated analysis is supposed to be evidence that *R* could be used as a statistical tool more widely, also by government administration or national research institutes.

References

EMA (2013). Basic data about electricity production, http://www.rynek-energii elektrycznej.cire.pl/st,33,207,tr,75,0,0,0,0,0,podstawowe-dane.html; [access in March, 2013]

Fott P. (1999). Carbon emission factors of coal and lignite: analysis of Czech coal data and comparison to European values, *Environmental Science & Policy*, 2 (1999) pp. 347–354

Kasztelewicz Z., Kozioł K. (2007). Production possibilities of brown coal industry in Poland after 2025. *Polityka Energetyczna*, vol.10, special issue 2, 2007 pp. 141–158 [in Polish]

Stefanović P., Marković Z., Bakić V., Cvetinović D., Spasojević V., Živković N. (2012a). Domestic Lignite Emission Factor Evaluation for Greenhouse Gases Inventory Preparation of Republic of Serbia, http://wbalkict.ipa.ro/wp-content/uploads/2012/06/Domestic-Lignite-Emission-Factor-Evaluation-for-Greenhouse-Gases-Inventory-Preparation-of-Republic-of-Serbia.pdf

Stefanović P., Marković Z., Bakić V., Cvetinović D., Spasojević V., Živković N. (2012b). Evaluation of Kolubara Lignite Carbon Emission Characteristics, *Thermal Science*, Year 2012, Vol.16, No. 3, pp. 805–816

Leveraging GPU libraries for efficient computation of Gaussian process models in *R*

Colin Rundel[1,⋆]

1. Duke University, Department of Statistical Science
⋆Contact author: rundel@gmail.com

Keywords: Gaussian processes, Rcpp, GPU, HPC, Armadillo

In this talk we will presents a case study of our recent work on the implementation of a Gaussian process based Bayesian model for spatial assignment. The talk will focus on the low level implementation of this model in *R* using **Rcpp**, **Armadillo**, and GPU linear algebra libraries **CUBLAS** and **Magma**. Building our implementation on top of these existing libraries allows us to exploit the computational power of commodity GPU hardware without the need for specific expertise in developing for these processors. We will discuss how through judicious use of these tools we are able to improve the performance of our assignment models by 3-5x over our original **RcppArmadillo** implementation. It is our hope that our case study will provide insight into the identification of common computational bottle necks which can be improved through the use of existing GPU libraries and implementations.

TriMatch: An R Package for Propensity Score Matching of Non-Binary Treatments

Jason M. Bryer[1,2,⋆]
Kimberly K. Speerschneider[1,2]

1. University at Albany
2. Excelsior College
⋆Contact author: jason@bryer.org

Keywords: propensity score analysis, matching, non-binary treatments

The use of propensity score methods (Rosenbaum and Rubin, 1983) have become popular for estimating causal inferences in observational studies in medical research (Austin, 2008) and in the social sciences (Thoemmes and Kim, 2011). In most cases however, the use of propensity score methods have been confined to a single treatment. Several researchers have suggested using propensity score methods with multiple control groups, or to simply perform two separate analyses, one between treatment one and the control and another between treatment two and control. This talk introduces the **TriMatch** package for R that provides a method for determining matched triplets. Examples from educational and medical contexts will be discussed.

Consider two treatments, Tr_1 and Tr_2, and a control, C. We estimate propensity scores with three separate logistic regression models where model one predicts Tr_1 with C, model two predicts Tr_2 with C, and model three predicts Tr_1 with Tr_2. The triangle plot in Figure 1 represents the fitted values (i.e. propensity scores) from the three models on each edge. Since each unit has a propensity score in two models, their scores are connected. The **TriMatch** algorithm will find matched triplets where the sum of the distances within each model is minimized. In Figure 1, the black lines illustrate one matched triplet.

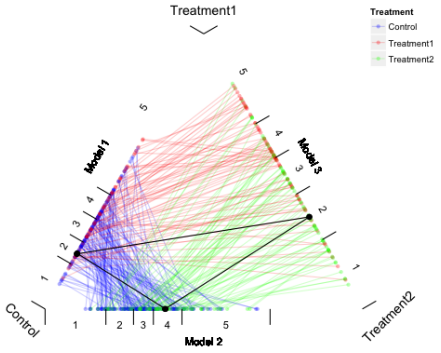


Figure 1: Triangle Plot

Propensity score analysis of two groups typically use dependent sample *t*-tests. The analogue for matched triplets include repeated measures ANOVA and the Freidman Rank Sum Test. The **TriMatch** package provides utility functions for conducting and visualizing these statistical tests. Moreover, a set of functions extending **PSAgraphics** (Helmreich and Pruzek, 2009) for matched triplets to check covariate balance are provided.

References

Austin, P. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine 27*, 2037–2049.

Helmreich, J. E. and R. M. Pruzek (2009, 2). Psagraphics: An r package to support propensity score analysis. *Journal of Statistical Software 29*(6), 1–23.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Thoemmes, F. J. and E. S. Kim (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research 46*, 90–118.

KmL3D: K-means for Joint Trajectories

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

1. UMR 1027, INSERM, Universit Paul Sabatier, Toulouse III, France
2. CeRSM (EA 2931), UFR STAPS, Universit de Paris Ouest-Nanterre-La Dfense, France
3. Research Unit on Children's Psychosocial Maladjustment, University of Montreal and Sainte-Justine Hospital, Montreal, Quebec, Canada
4. UI669 INSERM, Paris, France
5. University Paris-Sud and University Descartes, Paris, France

⋆Contact author: christophe.genolini@u-paris10.fr

Keywords: K-means, joint longitudinal data, cluster, graphical interface.

KmL3D is an implementation of the K-means algorithm specificaly design to cluster joint Longitudinal data.

Cohort studies are becoming essential tools in epidemiological research. In these studies, measurements are no longer restricted to a single variable but can be seen as variable-trajectory (for example "Evolution of stress"). When multiple variable-trajectories are considered simultaneously, they are called "joint trajectories" (example: "Joint evolution of stress and aggressiveness"). K-means is a statistical methods that can be used to determines homogeneous groups of patients joint trajectories.

KmL3D (Genolini et al., 2012) is an implementation of k-means design to work specifically with joint trajectory. Like **KmL** (see Genolini and Falissard (2010, 2011)), it provides friendly user graphical interface and facilities to deal with missing values. It can display the joint trajectories in 3D (either all the trajectories or the clusters mean) giving to the user an easy way to "see" the spatial shape of the clusters. It also provides some functions to export 3D graphes in pdf format (see figure 1). **KmL3D** can also work in higher dimension. The 3D graphical representation is then restricted at two variable-trajectories.

Figure 1: 3D rotating figure. Grab the graph with the left bouton and move the mouse to change the view.

References

Genolini, C. and B. Falissard (2010). KmL: k-means for longitudinal data. *Computational Statistics 25*(2).

Genolini, C. and B. Falissard (2011). KmL: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine 104*(3), 112–121.

Genolini, C., J. Pingault, T. Driss, S. Côté, R. Tremblay, F. Vitaro, C. Arnaud, and B. Falissard (2012). Kml3d: A non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine 109*, 104–111.

The objectives of this thesis was to model seasonal variations in claim intensities and to evaluate the dependency of covariates on claim rates. The data for this thesis were obtained from claimants registered during September 2009 to August 2011, both inclusive at the Ethiopian Insurance Corporation in Hawassa. We present a procedure for consistent estimation of the claim frequency for motor vehicles in the Ethiopian Insurance Corporation, Hawassa District. The seasonal variation is modeled with a non-homogeneous Poisson process with a time varying intensity function. Covariates of the policy holders, like gender and age, is corrected for in the average claim rate by Poisson regression in a GLM setting. An approximate maximum likelihood criterion is used when estimating the model parameters. The seasonal parameters are found to be $\pm 25\%$ and to be statistically significant. February has highest while August has lowest claim rate. Only age group 36-45 has significantly lower claim rate than age group 20-25. The rate is about one third. Lastly female is not found to have significantly lower claim rates than males, however, there are indications that maight be slightly so.

Introducing SimpleDataManager
A simple data management workflow for R

Finn Sandø[1,*], Camilla Trab Damsgaard[1], Stine-Mathilde Dalskov[1], Christian Ritz[1]

1. Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Denmark
*Contact author: fsp@life.ku.dk

Keywords: data management workflow, data formats, data hub structure, data storage

The aim of **SimpleDataManager** is to facilitate a data management workflow in a simple, easy to use fashion. Therefore **SimpleDataManager** does not intend to support analytical features except for some very simple validation functionality. The motivation behind **SimpleDataManager** is that many projects have a specific needs for data management that are better dealt with as separate tasks not tightly integrated into the subsequent analytical process. There have over the years been a few attempts at making such integrated workflow platforms into *R* packages. Some of these come with a graphical user interface, most notably John Fox' **Rcmdr**. A few other packages are currently available in the CRAN repository: **Rz** (a "GUI Tool for Data Management like SPSS or Stata") and **memisc** ("Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation"). These packages support data management, but are more focused on analytics and graphs. Some other data management initiatives have been developed within very specific contexts like **eiPack** ("Ecological Inference and Higher-Dimension Data Management").

Regarding ‘raw’ data management, a typical workflow could be as follows: Data are generated or obtained by one or more individuals from one or more sources. These data are collected centrally, some kind of clean-up, validation, and standardization steps are performed, and the data are merged into one or a few related datasets. An official standard for long-term storage may apply. Finally, there may be restrictions limiting who has access to which parts of the data, and usually there is a need to deliver in specific formats (*spss*, *sas*, *stata*, *csv*).

SimpleDataManager supports such a workflow. On the input side it operates with a repository of incoming ‘raw’ data, it promotes a unified and repeatable workflow for cleaning, validating, transforming, and merging data into a consistent format. On the output side it uses a request mechanism where recipients select required variables on an auto-generated list. Variables can be grouped to simplify the selection process. It supports missing values, multi-lingual variable and value labels defined by the user. Data can be exported to common formats (*sas*, *spss*, *stata*, *csv*), new formats can relatively easily be implemented. We are currently working on an extension that will make it possible to deliver data as a single *RData* file which will come with built-in functionality for export to all supported formats.

SimpleDataManager follows the *configuration by convention* pattern where, if the user adheres to a specific structure, most things will work 'automagically', it is even possible to configure the standard behavior with configuration files. **SimpleDataManager** will be made available through one of the public *R* repositories.

SimpleDataManager is developed as part of the large-scale research project OPUS (Optimal well-being, development and health for Danish children through a healthy New Nordic Diet), supported by a grant from the Nordea Foundation”.

SenseLabOnline: Combining agile data base administration with strong data analysis

Guillaume Le Ray[1*], Junaid Khalid[1]

1. DELTA, SenseLab
*Contact author: glr@delta.dk

Keywords: automated analysis, server-R communication, data base, sensory analysis

SenseLabOnline is a web application working as an online listening test facility. Concretely, SenseLabOnline enables the creation and administration of listening tests as well as panels. Assessors can be invited to a test that they can complete over the internet. The resulting data are then stored in a data base for ulterior analysis. SenseLabOnline is agile along the different steps, from the experimental design to the automatic statistical analysis through the Presentation of test stimuli. It has been developed based on various standards from the International Telecommunication Union.

SenseLabOnline combines the agility of a data base and the strong statistical functionalities of *R* packages like **car, FactoMineR, smacof**. It also relies on a range of graphical packages (**Hmisc, jpeg, gplots, EBImage**) to generate the output.

An XML file is exported from the data base with a predefined data structure containing all the test data. The *R*-coded routine imports that file using the **XML** package provided by *R*. The routine has been coded in S4, and it has been implemented in "blocs" of methods, which apply to a related class of objects and compiled into a package that is deployed on the server.

The collection of data into a data base and the analysis of sensory data using *R* have been used separately throughout the years. However SenselabOnline combines these two tools behind a common user interface, making it an original product.

References

Ramsgaard, J. Zacharov, N., Khalid, J., Le Ray, G. (2011). Evaluation of complex audio systems using Rapid Online Listening Test method, www.senselabonline.com, *Pangborn 2011 Symposium (Toronto, Canada)*

ITU-R (2003), Method for the subjective assessment of intermediate quality level of coding systems, *RECOMMENDATION ITU-R BS.1534-1*

ITU-R(1997), Methods foe the subjective assesment of small impairments in audio systems including multichannel sound systems, *RECOMMENDATION ITU-R BS.1116-1**

ITU-T (2003), Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, *RECOMMANDATION ITU-T P.835*

ffbase: statistical functions for large datasets

Edwin de Jonge[1], **Jan Wijffels**[2]

1. Statistics Netherlands, e.dejonge@cbs.nl
2. BNOSAC - Belgium Network of Open Source Analytical Consultants, jwijffels@bnosac.be

Keywords: Large datasets, memory constraints, modelling on large data

Statistical datasets used to be small, but nowadays it is not uncommon that the dataset is too large to be handled in R without encountering the frequently encountered **Error: cannot allocate vector of size ... Mb** issue.

To handle the *R* memory constraints, the **ff** package (Adler & Oehlschlägel et al.) was developed in 2008. It handles the memory constraint by storing data on disk. For day-to-day data munging, frequently used functionality from the **base** package had to be developed to make it more easy for an *R* developer to work with package **ff**. For this, the **ffbase** package has been developed to extend the **ff** package to allow basic statistical operations on large data frames, especially *ffdf* objects.

The **ffbase** package contains a lot of the functionality from the R's base package for usage with large datasets through package **ff**. Namely

- Basic operations (c, unique, duplicated, ffmatch, ffdfmatch, %in%, is.na, all, any, cut, ffwhich, ffappend, ffdfappend, rbind, ffifelse, ffseq, ffrep.int, ffseq_len)
- Standard operators $(+, -, *, /, \hat{}, \%\%, \%/\%, ==, != , <, \leq, \geq, >, \&, |, !)$ working on ff vectors
- Math operators (abs, sign, sqrt, ceiling, floor, trunc, round, signif, log, log10, log2, log1p, exp, expm1, acos, acosh, asin, asinh, atan, atanh, cos, cosh, sin, sinh, tan, tanh, gamma, lgamma, digamma, trigamma)
- Selections & data manipulations (subset, transform, with, within, ffwhich)
- Summary statistics (sum, min, max, range, quantile, hist, binned_sum, binned_tabulate)
- Data transformations (cumsum, cumprod, cummin, cummax, table.ff, tabulate.ff, merge, ffdfdply, as.Date, format)
- Chunked functionalities (chunkify), writing & loading data (load.ffdf, save.ffdf, move.ffdf, laf_to_ffdf)

For modelling purposes, **ffbase** has bigglm.ffdf to allow to build generalized linear models easily on large data and can connect to the **stream** package for clustering & classification.

In the presentation, the **ffbase** package will be showcased to show that working with large datasets without having RAM issues in *R* is easy and natural for an *R* programmer.

References

Daniel Adler, Christian Glser, Oleg Nenadic, Jens Oehlschlägel and Walter Zucchini (2013). ff: memory-efficient storage of large data on disk and fast access functions. R package version 2.2-11. http://CRAN.R-project.org/package=ff

Edwin de Jonge, Jan Wijffels and Jan van der Laan (2011). ffbase: Basic statistical functions for package ff. R package version 0.7-1. http://github.com/edwindj/ffbase

cold: a package for Count Longitudinal Data

M. Helena Gonçalves[1,2,⋆], **M. Salomé Cabral**[1,3]

1. Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
2. Departamento de Matemática, FCT, Universidade do Algarve, Portugal
3. Departamento de Estatística e Investigação Operacional, FCUL, Lisboa, Portugal

⋆ Contact author: mhgoncal@ualg.pt

Keywords: count longitudinal data, exact likelihood, first order autoregressive, random effects.

The software tools that we propose were built for the analysis of longitudinal count data from the point of view of likelihood inference, which requires complete specification of a stochastic model for the individual profile. We provide a direct evolution of the model proposed by Azzalini (1994) for Poisson response variables where the serial dependence is assumed to be of Markovian type. Besides serial dependence, another important source of dependence among data from one given subject is the presence of individual random effects (Gonçalves et al., 2007). The Poisson regression model which links the covariates and the probability distribution of the response, is $\ln\{E(Y_{it})\} = \ln(\theta_{it}) = x_{it}^{\top}\beta$ allowing also some form of dependence among observations of the same individual. The introduction of random effects can be formulated by adding the component $b_i \sim N(0, \sigma_b^2)$ in the previous model, leading to the random intercept model $\theta_{it}^b = \exp(x_{it}^{\top}\beta + b_i)$.

This software allows the presence of individual random effects by adding the component to the linear predictor. One dimensional integrals were computed using adaptive Gaussian quadrature. The package, called **cold**, is a S4-methods package and provides *R* functions for parametric analysis of longitudinal count data. The functions of **cold** were written in *R* language, except for some *FORTRAN* routines which are interfaced through *R*. The main function, called `cold` performs the fit of parametric models via likelihood methods.

Serial dependence and random effects are allowed according to the stochastic model chosen: `ind` (independence), `AR1` (1st order autoregressive), `indR` (independence with random effects), `AR1R` (1st order autoregressive with random effects). Missing values and unbalanced data are automatically accounted for computing the appropriate likelihood function.

References

Azzalini, A. (1994). Logistic regression and other discrete data models for serially correlated observations. *J. Ital. Statist. Soc. 2*, 169–179.

Gonçalves, M. H., M. S. Cabral, M. C. R. de Villa, E. Escrich, and M. Solanas (2007). Likelihood approach for count data in longitudinal experiments. *Computational Statistics and Data Analysis 51*, 6511–6520.

kPop: An *R* package for the interval estimation of the mean of the selected populations

Vik Gopal[1,⋆], **Claudio Fuentes**[2]

1. IBM Research Collaboratory, Singapore
2. Department of Statistics, Oregon State University
⋆Contact author: viknesh.gopal@gmail.com

Keywords: Confidence intervals, selected mean, selected populations, order statistics.

Consider the oneway balanced ANOVA experiment, where there are p populations, each with mean θ_i. Suppose that we can obtain a sample of size n from each of these populations, and denote each sample by X_{ij}, where $1 \leq i \leq p$ and $1 \leq j \leq n$. The standard ANOVA assumption is that $X_{ij} \sim N(\theta_i, \sigma^2)$. Under this distributional assumption, the task of prime interest in this paper is to estimate the confidence interval of the mean of the k populations that returned the largest sample averages. To be more precise, if $k = 1$, then we are interested in frequentist confidence intervals for the population mean θ_ℓ, chosen upon the criteria that the corresponding sample mean satisfies that $\bar{X}_{\ell\bullet} \geq \bar{X}_{j\bullet}$ for every population $j \neq \ell$.

Traditional intervals fail to maintain the nominal coverage probability as these methods are based on biased estimates and ignore the selection process. This bias becomes more pronounced as the means of the individual populations differ less. Consequently, it is important to take the selection mechanism into account in order to make correct inference.

Nowadays, this type of problems has become more relevant as researchers focus their attention on a subset of the populations under investigation. For instance, in genomic studies, researchers might be interested in identifying genes that are associated with a particular phenotype. In this context, typically thousands of genes are screened, but only a few are selected, using mechanisms such as false discovery rate(among others). Hence the inference on these selected genes has to be formally correct; otherwise the confidence coefficient would be much lower than the nominal level. In fact, the large number of 'populations' in the genetics setting accelerates the rate at which this confidence coefficient approaches 0.

In the *R* package **kPop**, we provide a collection of functions for experimenters to obtain confidence intervals, using different methodologies, for the selected population. Apart from traditional and Bonferroni intervals, **kPop** provides the following intervals:

1. When $k = 1$, it is straightforward to derive symmetric confidence intervals with the desired coverage probability. The package extends this approach by including a procedure for reducing the length of the interval when the experimenter has some knowledge pertaining to the difference between the means. This approach is explained in Fuentes and Casella (2012). In addition, **kPop** provides bootstrap intervals for the selected population when $k = 1$, using w-estimators, which correct for the bias of selection while optimizing the mean square error. This general class of estimators are covered in Venter and Steel (1991) and basically correspond to weighted combinations of the order statistics of the sample means.

2. The main contribution of the package, however, is when $k > 1$. In this case, the approach in Fuentes and Casella (2012) for $k = 1$ can be generalised, thus providing a practical yet formal tool for estimating (simultaneously) the mean of several populations.

In addition, the package provides plotting functions for visualizing the different intervals. In the talk, we shall give a summary of the above techniques and their application to a dataset, which shall be included in the package **kPop**.

GLM - a case study: Antagonistic relationships between fungi and nematodes

Tatjana Kecojević[1⋆]

1. University of Central Lancashire, UK
⋆Contact author: tkecojevic@uclan.ac.uk

Keywords: generalised linear model, quasi-binomial, logit

The aim of this study was to confirm the influence of a particular set of fungi on the scion of nematodes. The seasonally collected data of the proportions of successful scion of the nematodes influenced by the selected fungi provided for this study comprised of only four replicates using six different fungi. Considering that the response was a continuous variable in the interval [0, 1] meant that it could be analysed as a quasi-binomial variable using a generalised linear model (GLM) [1]. In order to confirm the evidence of a practically apparent relationship using such a small data set the response was treated as a specific quasi-binomial variable with a logit link and variance function of the form $\mu^2(1-\mu)^2$ [2].

The talk illustrates the analysis of the problem and shows how the family argument of the `glm` function in *R* had to be expanded in order to enable the application of this specific quasi-GLM that accommodates this particular link and variance function.

References

[1] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

[2] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika 61*(3), 439–447.

R Packages for Rank-based Estimates

John Kloke[1,⋆], **Joseph McKean**[2]

1. University of Wisconsin-Madison, Madison, WI
2. Western Michigan University, Kalamazoo, MI
⋆Contact author: kloke@biostat.wisc.edu

Keywords: Robust, Linear Models, Cluster Correlated

Rank-based estimates are a class of robust estimates for which the objective function is a function of the ranks. Inference and diagnostic procedures for a linear model assuming iid errors are well established (see for example McKean & Hettmansperger 2011) and implemented in the *R* package **Rfit** (Kloke, McKean 2012). Methods which extend rank-based estimates to include cluster correlated data are in development. One such approach is the joint-rankings estimate (Kloke, McKean, Rashid 2009). We are in the process of developing an R package for these joint-ranking estimates **jrfit** which extends **Rfit**.

In this talk, we briefly summarize the theory for rank and joint-ranking estimates. For the majority of the talk, however, we present a series of examples illustrating the use of **Rfit** and **jrfit**.

References

Kloke, JD, McKean, JW (2012). Rfit: Rank-based Estimation for Linear Models. *R Journal*, 4/2, 57-64.

Kloke, JD, McKean, JW, Rashid, M (2009). Rank-Based Estimation and Associated Inferences for Linear Models with Cluster Correlated Errors. *Jounral of the American Statistical Association*, 104, 384-390.

McKean, JW, Hettmansperger, TP (2011). *Robust Nonparametric Statistical Methods, Second Edition*, Boca Raton, FL:Chapman Hall.

Heart Rate Variability analysis in R with RHRV

Constantino A. García[1,⋆], Abraham Otero[2], Jesús Presedo[1], Xosé A. Vila[3]

1. Centro Singular de Investigación en Tecnoloxías da Información (CITIUS), University of Santiago de Compostela, Santiago de Compostela, Spain.
2. Department of Information and Communications Systems Engineering, University San Pablo CEU, 28668, Madrid, Spain.
3. Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004, Ourense, Spain.
⋆Contact author: constantinoantonio.garcia@usc.es

Keywords: Heart Rate Variability, time-domain analysis, frequency-domain analysis, nonlinear analysis.

Abstract

Heart Rate Variability (HRV) refers to the variation over time of the intervals between consecutive heartbeats. Since the heart rhythm is modulated by the autonomic nervous system (ANS), HRV is thought to reflect the activity of the sympathetic and parasympathetic branches of the ANS. The continuous modulation of the ANS results in continuous variations in heart rate. HRV has been recognized to be a useful non-invasive predictor of several pathologies such as myocardial infarction, diabetic neuropathy, sudden cardiac death and ischemia, among others [2] and it is an active research field with hundreds of publications every year.

These publications are often hard to replicate since their authors have implemented their own analysis algorithms. Furthermore, the implementation of these algorithms is time consuming and prone to error. This situation also often prevents clinicians and researchers from following the state-of-the-art HRV analysis methods (such as the use of wavelet based HRV frequency analysis) due to the complexity of implementing the analysis algorithms themselves. We believe that freely available software tools including the most common HRV analysis algorithms may alleviate this problem, and it may speed up advances in HRV research.

For these reasons we have developed **RHRV**, an opensource package for the *R* environment that comprises a complete set of tools for HRV analysis [1], [3]. **RHRV** allows the user to import data files containing heartbeat positions in the most broadly used formats; eliminating outliers or spurious points present in the time series with unacceptable physiological values; plotting HRV data and performing time domain, frequency domain and nonlinear HRV analysis. **RHRV** is the first *R* package for HRV analysis.

By using **RHRV**, the authors would not have to implement the most broadly used analysis algorithms themselves and they shall have at their disposal the statistical and numerical analysis power of *R* for analyzing HRV data. Thus, **RHRV** provides interesting functionality for clinicians and researchers with an interest on systematizing HRV analysis or developing new HRV analysis algorithms.

References

[1] García, C., A. Otero, X. Vila, and M. Lado (2012). An open source tool for heart rate variability wavelet-based spectral analysis. In *International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSIGNALS 2012*.

[2] Kautzner, J. and A. John Camm (1997). Clinical relevance of heart rate variability. *Clinical cardiology 20*(2), 162–168.

[3] Rodríguez-Liñares, L., A. Méndez, M. Lado, D. Olivieri, X. Vila, and I. Gómez-Conde (2010). An open source tool for heart rate variability spectral analysis. *Computer Methods and Programs in Biomedicine*.

Massively Parallel Computation of Climate Extremes Indices using R

David Bronaugh[1,*], Jana Sillmann[2]

1. Pacific Climate Impacts Consortium
2. Canadian Centre for Climate Modeling and Analysis
*Contact author: bronaugh@uvic.ca

Keywords: parallel, climate, extremes, performance

The **climdex.pcic** package provides an R implementation of the Expert Team on Climate Change Detection and Indices' CLIMDEX climate extremes indices. It has been used to compute these indices on a global scale for hundreds of model runs from dozens of different climate models, totaling terabytes of data. Performance using **climdex.pcic** is comparable to the *FORTRAN* implementation (fclimdex) which preceded it. Because it is written in *R*, parameterization and automation became trivial. Furthermore, parallelizing the code using the **snow** package was trivial and provided a near linear speedup with increasing processor cores. Combining this with custom job dispatch code enabled the use of the WESTGRID cluster, decreasing compute time from months under fclimdex to days under **climdex.pcic**.

The **climdex.pcic** package has been released on CRAN; the supporting code to compute indices in parallel on a cluster will be released on CRAN in the near future. The code is currently available on the Pacific Climate Impacts Consortium's website.

References

Bronaugh, D. (2013). CRAN: package climdex.pcic, http://pacificclimate.org/resources/software-library

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multi-model ense.mble. Part 1: Model evaluation in the present climate. *Journal of Geophysical Research – Atmospheres,* doi:10.1002/jgrd.50203, in press.

Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., & Bronaugh, D. (2013). Climate extreme indices in the CMIP5 multi-model ensemble. Part 2: Future climate projections. *Journal of Geophysical Research – Atmospheres,* doi:10.1002/jgrd.50188, in press.

Expert Team on Climate Change Detection and Indices (2001). Definitions of the 27 core indices. http://etccdi.pacificclimate.org/list_27_indices.shtml

Segmentor3IsBack: an *R* package for the fast and exact segmentation of Seq-data

Alice Cleynen[1,2,⋆], **Michel Koskas** [1,2], **Emilie Lebarbier** [1,2], **Guillem Rigaill**[3], **Stéphane Robin** [1,2]

1. AgroParisTech, UMR 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
2. INRA, UMR 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
3. Unité de Recherche en Génomique Végétale (URGV) INRA-CNRS-Université dEvry Val dEssonne, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France.

⋆Contact author: alice.cleynen@agroparistech.fr

Keywords: Segmentation, Negative Binomial, Seq-data, Fast, Exact

The concept of segmentation is simple: identify contiguous segments in the data with similar characteristics (typically shared parameter of the model distribution) in order to discover significant regions. The output of Next-Generation Sequencing experiments is a typical dataset where segmentation can proove useful: for instance we can identify portions of a chromosome with identical copy numbers, with a resolution unequaled by the previous CGH-arrays technologies. Another promising example is the identification and localization of transcribed regions of the genome through the segmentation of poly(A) RNA-Seq data: indeed, exons of a given genome are separated by intronic or non-coding regions where the signal is expected to reveal low transcriptional activity.

Our algorithm, the Pruned Dynamic Programming Algorithm (PDPA) and its implementation in the *R* package **Segmentor3IsBack** (available on the CRAN) for the negative binomial distribution addresses these questions while overcoming two major difficulties: the important size of the profiles (up to 10^9 data-points), and the discrete nature of the output (number of reads starting at each position of the genome).

The function `Segmentor` returns the optimal segmentation of the data in 1 to K_{max} segments with respect to the likelihood criterion with a complexity empirically faster than $\mathcal{O}(K_{max} n \log n)$ for distributions including the negative binomial (adapted to Seq-data analysis), Poisson, or Gaussian homoscedastic. Depending on the chromosome size and the choosen K_{max}, its runtime for the analysis of a real data-set varies between 20 minutes and 1 hour. The function `SelectModel` then allows the user to select the number of segments according to various penalty criteria including the standard BIC and its modified version mBIC (Zhang and Siegmund, 2007), and oracle penalties developed for each distribution (Lebarbier, 2005; Cleynen and Lebarbier, 2013). We investigated the advantage of the negative binomial model over the Poisson model using real RNA-Seq data and resampled RNA-seq data.

References

Cleynen, A. and E. Lebarbier (2013). Segmentation of the poisson and negative binomial rate models: a penalized estimator. Arxiv preprint arXiv:1301.2534.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing 85*(4), 717–736.

Zhang, N. R. and D. O. Siegmund (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics 63*(1), 22–32.

hts: R tools for hierarchical time series

Rob J Hyndman

Department of Econometrics and Business Statistics, Monash University, Australia
rob.hyndman@monash.edu

Keywords: forecasting, GLS regression, R package, reconciling forecasts, time series.

Hierarchical time series occur when there are multiple time series that are hierarchically organized and can be aggregated at several different levels in groups based on dimensions such as product, geography, or some other features. A common application occurs in manufacturing where forecasts of sales need to be made for a range of different products in different locations. The forecasts need to add up appropriately across the levels of the hierarchy.

I will describe the **hts** package for *R* which provides several methods for analysing and forecasting hierarchical time series. These include the popular "bottom-up" method, various "top-down" methods, "middle-out" approaches, as well as the optimal combination approach based on a large ill-conditioned regression model.

Research issues include how to generate prediction intervals that account for the hierarchical structure, and the use of various computational tricks in order to implement the forecasting methods in a realistic time frame for operational purposes.

Teaching statistics interactively with Geogebra and R

Virgilio Gómez-Rubio[1,⋆], María José Haro-Delicado[1] and Francisco Parreño-Torres[1]

1. Department of Mathematics, Universidad de CAstilla-La Mancha (Spain)
⋆Contact author: virgilio.gomez@uclm.es

Keywords: Geogebra, R, Teaching Statistics, Interactive Teaching

Geogebra (Geogebra Team, 2013) is a software for dynamic mathematics and geometry. With Geogebra, students can create and manipulate geometric objects in an interactive way. A simple example in Geogebra is to create to points and the line that they define. Then, by clicking and moving one of the points the line will de updated accordingly. More complex examples can be created by using other features available in **Geogebra**. This software provide a valuable resource to teach mathematics (and geometry, in particular) but it could also be used to teach statistics.

In this work, we describe how to link **Geogebra** and *R* (R Core Team, 2013) using package **Rserve** (Urbanek, 2013). **Geogebra** is implemented in *JAVA* but it provides the possibility of developing scripts in other languages, such as *Python*. **Geogebra** provides a "*Python* window" (based on *Jython*, an implementation of *Python* using *JAVA*) to develop small scripts. These scripts can be linked to actions on the geometric objects so that, for example, the script is run when a point is clicked on or moved.

Initially, this was done by loading the **Rserve** API (in a JAR file) from the *Python* window but, as this API is also written in *JAVA*, it was later added to the **Geogebra** source code. Developing examples will require the use of *Python* and embedded *R* code. As difficult as it may sound, we will show how simple and complex examples can be easily developed. The main advantage of having this link is that **Geogebra** can use the myriad of statistical methods implemented in *R* in a very easy way.

In particular, we will consider an example on quality control where control bands change automatically when the user changes the observed points. We will also show to import maps in **Geogebra** and other examples in spatial statistics. In all these examples, users can manipulate the observed data so that the results are automatically recomputed and displayed when these objects are changed. This provides a unique environment to teach probability and statistics interactively.

References

Geogebra Team (2013). Geogebra: Dynamic mathematics & science for learning and teaching. http://www.geogebra.org/.

R Core Team (2013). R: A language and environment for statistical computing. http://www.R-project.org/. ISBN 3-900051-07-0.

Urbanek, S. (2013). Rserve: Binary r server. http://www.rforge.net/Rserve/. R package version 1.7-0.

RKTeaching: A new R package for teaching Statistics

Alfredo Sánchez-Alberca[1,⋆]

1. San Pablo CEU University
⋆Contact author: asalber@ceu.es

Keywords: RKWard, RKTeaching, Teaching, Graphical User Interface.

Step by step, *R* is becoming one of the main software used by the scientific community for data analysis, displacing other giants like *SPSS*, *SAS* or *STATA*. *R* has a lot of strengths, most of them as a consequence of being open source, but its main weakness is the lack of a mature Graphical User Interface (GUI), making it a little intimidating for students and beginners. That is the main reason that R has not spread yet to other public, especially in the field of education. Fortunately, in the last years some GUI are emerging to overcome this drawback, as for example **R commander**, **JGR**, **RStudio** or **RKWard**, but they are still not as user friendly as commercial GUIs.

One of the most promising GUI is **RKWard** [Rödiger 2012]. RKWard has a lot of advantages over its competitors. First, it is open source and multi platform. Second, it is based in KDE and QT graphics libraries, much more modern than the tcl/tk libraries that uses **R commander**. And third, the most important, unlike **RStudio**, it is highly expandable and customizable by means of plugins. In addition, **RKWard** is not just a GUI, but a complete development environment that has been designed both for beginners and experts.

For all these reasons, I decided three years ago to develop a new *R* package for teaching statistics based on **RKWard**. My main goal was to simplify and facilitate the use of *R* to reduce the learning curve. The new package is called **RKTeaching** and, at this moment, it includes menus, dialogs and procedures for data preprocessing (filtering, recoding, weighting), frequency tabulation (grouped and non grouped data), plotting (bar charts, pie charts, histograms, box charts, scatter plots), descriptive statistics (mean, median, mode, percentiles, variance, unbiased variance, std deviation, unbiased std deviation, coefficient of variation, intercuartile range, skewness coefficient, kurtosis coefficient), probability distributions (Binomial, Poison, Uniform, Normal, Chi square, Student t, Fisher's F, exponential), parametric tests (T test for one mean and mean comparison (independent and paired samples), F test for variances comparison, Normal and Binomial tests for one proportion and proportions comparison, ANOVA for multiple factors and repeated measures, sample size calculation for estimating means and proportions), non parametric tests (Kolmogorov-Smirnov, Shapiro Wilks, Mann Whitney U, Wilcoxon, Kruskal Wallis, Friedman, Chi square), concordance tests (intraclass correlation coefficient, Cohen's Kappa), regression and correlation (linear, non linear, logistic, models comparison, predictions), simulations (coin tosses, dice roll, small numbers law, central limit theorem). All these menus include optional assistance that guide the user step by step trough the analysis and help interpret the results. In addition, some procedures have an extended version where all the calculations and formulas used in the analysis are displayed in order to help students to understand the procedure.

In the last two years we have used this package to teach statistics to Medical, Pharmacy, Psychology, Biotechnology, Optics, Nursing and Nutritional sciences degrees, with very satisfying results. To assess the achievements we conducted an experiment to compare **RKTeaching** with previous software. The assessment by the students clearly reflects its ease of use and learning compared to *SPSS* [Sánchez 2012].

References

Rödiger, S et al. (2012). RKWard: A Comprehensive Graphical User Interface and Integrated Development Environment for Statistical Analysis with R. *Journal of Statistical Software 49(9)*, 1–34.

Sánchez-Alberca, A. (2012). RKTeaching: un paquete de R para la enseanza de Estad´. In *III Jornadas de Intercambio de Experiencias de Innovacin Educativa en Estad´(Valencia, Spain)*, pp. 136–147.

genertest: a package for the developing exams in R

Lara Lusa[1,⋆]

1. ⋆Institute for Biostatistics and Medical Informatics, University of Ljubljana, Slovenia;
⋆Contact author: lara.lusa@mf.uni-lj.si

Keywords: Teaching, Sweave, RExcel

Preparing the tests for written examinations can be a very time-consuming task for instructors. Test preparation can be even more demanding if some stratagem has to be adopted to prevent students from cheating during written exams, for instance when multiple versions of same test are necessary.

Many instructors collect the past tests in electronic format and maintain an updated questions database. The availability of a questions database can considerably reduce the time needed for the preparation of tests for new examinations especially if some specialized software for the preparation of tests is used.

This paper presents **genertest**, an R package for development of tests. The questions to include in the test can be sampled based on their topic and/or on the number of points; the user can decide how many versions of the test should be produced, if the order of the questions and/or sub-questions should be permuted, etc. The tests can be customized providing the name of the course and date of the examination, and languages different than English can be used. The database of questions needed for using **genertest** is a tab-delimited text file and it has a simple structure. **Sweave** syntax can be used in the formulation of questions and answers, and the tests can contain pictures.

genertest can be invoked also from Microsoft ®Office Excel (Excel), where the end-user can specify how to generate the tests answering to twenty questions in "plain-English" contained in an Excel spreadsheet, and obtain the printable tests simply by invoking an Excel macro written in Visual Basic for Applications. To manage the connection between Excel and R we use the RExcel add-in for Excel together with the statconn server and rcom package, all of which were developed within the statconn project (Baier and Neuwirth, 2007).

The package is available at
http://sites.google.com/site/lara3107/Home/software/genertest.

References

Thomas Baier and Erich Neuwirth (2007). Excel :: Com :: R. *Computational Statistics, 22*, 91–108.

Flexible generation of e-learning exams in *R*: Moodle quizzes, OLAT assessments, and beyond

Achim Zeileis[1,*], **Nikolaus Umlauf**[1], **Friedrich Leisch**[2]

1. Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Austria
2. Institute of Applied Statistics and Computing, Universität für Bodenkultur Wien, Austria
*Contact author: Achim.Zeileis@R-project.org

Keywords: exams, e-learning, multiple choice, arithmetic problems, `Sweave`

The capabilities of the package **exams** for automatic generation of (statistical) exams in *R* are extended by adding support for learning management systems: As in earlier versions of the package, exam generation is based on separate `Sweave` files for each exercise – but rather than just producing different types of PDF output files, the package can now render the same exercises into a wide variety of output formats. These include HTML (with various options for displaying mathematical content) and XML specifications for online exams in learning management systems such as **Moodle** or **OLAT**. This flexibility is accomplished by a new modular and extensible design of the package that allows for reading all weaved exercises into *R* and managing associated supplementary files (such as graphics or data files). The readily available user interfaces are introduced with special emphasis on **Moodle/OLAT**. Furthermore, it is outlined how the underlying infrastructure is designed and how new functionality can be built on top of the existing tools.

References

Grün, B. and A. Zeileis (2009). Automatic generation of exams in *R*. *Journal of Statistical Software 29*(10), 1–14. http://www.jstatsoft.org/v29/i10/.

Leisch, F. (2002). Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), *COMPSTAT 2002 – Proceedings in Computational Statistics*, Heidelberg, pp. 575–580. Physica Verlag.

Zeileis, A., N. Umlauf, and F. Leisch (2012). Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond. Working Paper 2012-27, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck. http://EconPapers.RePEc.org/RePEc:inn:wpaper:2012-27.

Teaching R in the Cloud

Karim Chine[1,*]

1. Cloud Era Ltd
*Contact author: karim.chine@gmail.com

Keywords: e-Learning, distant education, Cloud Computing, EC2, Collaboration

Cloud Computing is holding the promise of democratizing access to computing infrastructures but the question "How will we bring the Infrastructure-as-a-Service paradigm to the statistician's desk and to the statistics classroom?" has remained partly unanswered. The Elastic-R Software platform proposes new concepts and frameworks to address this question: R, Python, Matlab, Mathematica, Spreadsheets, etc. are made accessible as articulated, programmable and collaborative components within a virtual and immersive education environment. Teachers can easily and autonomously prepare interactive custom learning environments and share them like documents in Google Docs. They can use them in the classroom or remotely in a distant learning context. They can also associate them with on-line-courses. Students are granted seamless access to pre-prepared, controlled and traceable learning environments. They can share their R sessions to solve problems in collaboration. Costs may be hidden to the students by allowing them to access temporarily shared institution-owned resources or using tokens that a teacher can generate using institutional cloud accounts.

References

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

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

[3] www.elastic-r.net

[4] www.coursera.org

[5] aws.amazon.com

BayesClass: An R package for learning Bayesian network classifiers

Bojan Mihaljevic[1,⋆], **Pedro Larrañaga**[1], **Concha Bielza**[1]

1. Computational Intelligence Group, Departamento de Inteligencia Artificial Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo sn, 28660, Boadilla del Monte, Madrid
⋆Contact author: b.mihaljevic@alumnos.upm.es

Keywords: Bayesian network classifiers, supervised classification, machine learning

BayesClass implements ten algorithms for learning Bayesian network classifiers from discrete data. This includes score+search algorithms and those that maintain the structure of a naive Bayes but extend it with additional parameters. Many of the algorithms perform implicit feature selection. Most of the structure searching methods are based on the forward search heuristic while score functions include classifier accuracy, likelihood, and a function based on the significance of dependency between two sets of variables (see Blanco et al. (2005)). Implemented algorithms include: *tree augmented naive Bayes* (TAN) (Friedman et al. (1997)), *forward sequential selection* (Langley and Sage (1994)), *forward sequential selection and joining* (FSSJ) (Pazzani (1996)), and adaptations of those methods from (Blanco et al. (2005)); the *adjusted probability naive Bayesian classification* (Webb and Pazzani (1998)), the *attribute weighted naive Bayes* (Hall (2007)), and others. We also propose an adaptation of the TAN algorithm to operate on structures learned by FSSJ.

The assessment and use of induced classifiers are straightforward. An interface to the **caret** package allows for estimation of predictive performance by resampling. Several discretization procedures are implemented. Discretization learned from training data can be applied to test data by mapping real-valued unseen data to the intervals learned, so that the effect of discretization on classifier performance can be assessed. All algorithms can handle incomplete training data, approximating the sufficient statistics of a probability function corresponding to some node X by ignoring cases with missing values for either X or any of its parents. An interface to the **gRain** package provides inference and its integration with the **Rgraphviz** package, thus enabling graph plotting. We expect to publish **BayesClass** on CRAN during april of 2013.

References

Blanco, R., I. Inza, M. Merino, J. Quiroga, and P. Larrañaga (2005). Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics 38*(5), 376–388.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning 29*, 131–163.

Hall, M. (2007). A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems 20*(2), 120–126.

Langley, P. and S. Sage (1994). Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence (UAI-1994)*, pp. 399–406. Morgan Kaufmann.

Pazzani, M. (1996). Constructive induction of cartesian product attributes. In *Proceedings of the Information, Statistics and Induction in Science Conference (ISIS-1996)*, pp. 66–77.

Webb, G. I. and M. J. Pazzani (1998). Adjusted probability naïve Bayesian induction. In *Proceedings of the 11th Australian Joint Conference on Artificial Intelligence (AI-1998). Lecture Notes in Computer Science*, Volume 1502. Springer.

Constructing fuzzy rule-based systems with the R package "frbs"

L.S. Riza*, C. Bergmeir, F. Herrera, J. M. Benítez

Dept. of Computer Science and Artificial Intelligence, DiCITS Lab, SCI2S group
CITIC-UGR, Universidad de Granada, Granada, Spain

*Contact author: lala.s.riza@decsai.ugr.es

Keywords: fuzzy rule-based systems, classification, regression

Fuzzy sets as proposed by Zadeh (1965) are a generalization of classical set theory, in which objects, instead of just being members of a set or not, have a gradual degree of membership. Fuzzy rule-based systems (FRBS) have been used in the past successfully in many applications. They are competitive methods for classification and regression, especially for complex problems. One of their leading properties is that they are usually easy to interpret.

In CRAN, there exist already some packages for building FRBSs. The **sets** package (Meyer and Hornik 2009) implements the fundamental operations on fuzzy sets, and allows to build Mamdani-type FRBSs manually. The package **fugeR** (Bujard 2012) implements a method that is capable of learning FRBSs from data using a coevolutionary genetic algorithm.

We present the **frbs** package (published on CRAN, Riza *et al.* 2013), which is focused on the deployment of FRBSs, and their construction from data using procedures from Computational Intelligence (e.g., neural networks and genetic algorithms) to tackle classification and regression problems. The types of FRBSs considered in the package are Mamdani, Takagi Sugeno Kang, and other variants. For the learning process, **frbs** provides a host of standard methods, such as Wang & Mendel's technique, ANFIS, HyFIS, DENFIS, substractive clustering, SLAVE, and several others. The package allows for a flexible model construction by implementing multiple choices for operators (conjunction, disjunction, implication, etc.) and membership functions (e.g., triangle, trapezoid, Gaussian, etc.). Additionally, our package also allows for constructing FRBSs from human expert knowledge. In sum, with the package **frbs** we present a package that provides the most widely used algorithms and a comprehensive methodology in the field of FRBSs for regression and classification.

Acknowledgements

This work was supported in part by the Spanish Ministry of Science and Innovation under Projects TIN2009-14575, TIN2011-28488, and P10-TIC-06858. Lala S. Riza is grateful to the Dept. of Computer Science, Universitas Pendidikan Indonesia, for support in pursuing the PhD program. C. Bergmeir holds an FPU scholarship from the Spanish Ministry of Education.

References

Bujard A (2012), fugeR: Fuzzy Genetic, a Machine Learning Algorithm to Construct Prediction Model Based on Fuzzy Logic, R package version 0.1.2, URL http://CRAN.R-project.org/package=fugeR.

Meyer D, Hornik K (2009), Generalized and Customizable Sets in R, Journal of Statistical Software, Vol. 31, No. 2.

Riza, L. S., C. Bergmeir, F. Herrera, and J. M. Benitez (2013), frbs: Fuzzy Rule-based Systems for Classification and Regression Tasks, R package version 2.1-0, http://CRAN.R-project.org/package=frbs.

Zadeh L (1965), Fuzzy Sets, Information and Control, Vol. 8, pp. 338 - 353.

bbRVM: an R package for Ensemble Classification Approaches of Relevance Vector Machines

Selçuk Korkmaz[1,*], Dinçer Göksülük[1], Gökmen Zararsız[1]

1. Hacettepe University Faculty of Medicine Department of Biostatistics
*Contact author: selcuk.korkmaz@hacettepe.edu.tr

Keywords: Relevance vector machines, boosting, bagging, ensemble classification

Relevance vector machines are kernel-based tools being explored for classification and regression problems in last few years. They use a probabilistic Bayesian learning framework for classification and has a number of advantages over the popular and state-of-the art tool support vector machines. During the last decade, ensemble methods have been proposed to achieve higher classification accuracies than the single methods by using multiple models and aggregating the results of each model. Boosting and bagging are the most common ensemble techniques and applied to many algorithms such as decision trees, perceptrons, k-nearest neighbor classifiers, etc. In this study, we will describe **bbRVM** package which implements the boosting and bagging ensembles of relevance vector machine classification. We will also give examples from real datasets to demonstrate the applicability of the package.

References

Tipping M.E. (2001) Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, 211-244.

Classification Using C5.0

Max Kuhn[1⋆]

1. Pfizer Global R&D, Groton, Connecticut, USA
⋆Contact author: mxkuhn@gmail.com

Keywords: Classification Tree, Rule–Based Model, Boosting, Feature Selection

C4.5 (Quinlan, 1993) is a classification methodology that was developed in the same era as the classification and regression tree (CART) methodology (Breiman et al., 1984) but is more widely known outside of statistics. The model can the the form of a tree–based model or a set of distinct rules. C5.0 is an evolved version of this algorithm and, although it is very similar to C4.5, very little has been published on the primary improvements: feature selection ("winnowing") and boosting. In 2010, Quinlan created an open–source version of the software, which we adapted into the **C50** package. This talk briefly describes in the improvements in C5.0 as well as the functionality available in the package.

References

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Extending the Reach of R to the Enterprise

Lou Bajuk-Yorgan[1*]

1. TIBCO Software
*Contact author: lbajuk@tibco.com

Keywords: R, Commercial, Performance

We have developed a brand-new, high-performance, *R*-compatible statistical engine: **TIBCO Enterprise Runtime for R (TERR)**. This was done in response to the challenges that we have heard from many customers, as they have sought to more widely leverage *R* in their enterprise. As the creators of **S-PLUS**, with a strong history of bringing the benefits of an *S* language engine to commercial organizations, we are uniquely suited for this development. The goal of this new engine is to enable *R* users to develop in open source *R*, and then to deploy, scale and integrate their scripts and packages using TERR, without having to change any of their *R* code.

In this talk, I will discuss these challenges, as well as our journey to developing **TERR**, from a philosophical, technical and intellectual property perspective. I will also discuss the benefits it brings to both large organizations and individual *R* users, examples of **TERR** usage, our relationship with the open source *R* community, and how *R* users can freely access **TERR**.

References

TIBCO Spotfire (2012). Announcement of Spotfire 5.0 and TERR http://spotfire.tibco.com/en/about-us/news-room/press-releases/2012/09-25-12-spotfire-5.aspx.

Bajuk-Yorgan, Louis J (2013). TERR Community Site. https://www.tibcommunity.com/community/products/analytics/terr

TIBCO Spotfire (2013). TERR Developer Edition Download. http://tap.tibco.com/storefront/trialware/tibco-enterprise-runtime-for-r/prod15307.html

Big-data, real-time R? Yes, you can.

David Smith[1]

1. Revolution Analytics
*Contact author: david@revolutionanalytics.com; @revodavid

Keywords: R, real-time, big data, deployment, Revolution Analytics

In the commercial sector, *R* is widely used as a prototyping environment in research and development departments. And given that companies in all industries are increasing their use of statistical models (aka "predictive analytics") in automated business processes[1], you'd think it would be a natural choice for the IT department to use *R* there as well. But at Revolution Analytics, we've sometimes had to overcome some preconceptions about *R* that have been a barrier to its use in real-time production environments. This purpose of this talk is to dispel those preconceptions. In this talk, I'll describe a production architecture around *R* that we've developed for Revolution R Enterprise. This architecture is used by our clients for estimating statistical models on big data stored in databases, data warehouses and Hadoop, and to deploy statistical models to real-time environments. During the talk I'll explain what the phrases "big data" and "real time" actually mean in this contexts, motivated by some examples of companies who have successfully deployed *R* in real-world big-data, real-time applications.

References

1. McKinsey Global Institute (2011), "Big data: The next frontier for innovation, competition, and productivity", http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Large-Scale Predictive Modeling with R and Apache Hive: from Modeling to Production

Alex Zolotovitski[1,*] and Yakov Keselman[1]

1. Medio Systems Inc www.medio.com
*Contact author: alex@zolot.us

Keywords: HPC, hive, hadoop streaming, deployment of R models in production

Some suggestions:

Recently, we have been building predictive models on large (e.g., 200M users generating 50B events) HDFS data sets related to user behavior. We have experimented with several existing frameworks (such as *Apache Mahout*) and *R* packages (see, e.g., those listed in *Hadoop* section of HPC task view on CRAN [1]) that are able to build models based on full data sets. We have observed that while the final models were quite accurate, the time to iterate on model building and model validation was exceedingly high.

To significantly reduce the time it takes to build and validate a predictive model, we experimented with an alternative approach that uses local *R* on researcher's laptop as a client for *hive server* pulling samples of data for fast and rich graphics, descriptive statistics etc. and for building nearly as accurate models on samples of the full data set. Moreover, the resulting *R* code, with minimal changes (e.g., redirecting graphical and diagnostic output, read data from the standard input, output results to standard output), is suitable for processing the full data set via *Hadoop/Hive streaming* (after ensuring that all data for the same user is chunked together and ordered by time) and *Rscript*. Hence, validation of the best locally-developed models on the full data set is fast as well.

In our presentation, we illustrate the full cycle of predictive modeling on large mobile gaming data sets. First, Hive is used for sampling (stratified, if necessary) of user data and for defining simple cumulative attributes on the data. Second, *R* is used for fast iterative predictive model building and validation on the sample and for defining additional attributes, if needed. Next, the resulting model is executed on the full data set by streaming user data (one user at a time) through the *R* scripts that were produced during the modeling step. Finally, performance of the model is monitored by executing additional *R* scripts on a *Hive* sample of the full data set. Our experience shows that such arrangement takes full advantage of both tools, resulting in accurate models that scale to hundreds millions of users.

References

1. High Performance Computing CRAN Task View http://cran.r-project.org/web/views/HighPerformanceComputing.html .

Non-Life Insurance Pricing using R

Allan Engelhardt[1,*] , Suresh Gangam[2]

1. CYBAEA Limited, London, England
2. 64 Squares, Maharashtra, India
*Contact author: Allan.Engelhardt@cybaea.net

Keywords: Insurance, GLM, Ensemble models

Insurance have greatly benefited from adopting the *R* platform and leading companies are already reaping the rewards. We will show one example from non-life insurance pricing which will cover both technical implementation and business change, and we will share information on the commercial benefits obtained. By using a specific example we can keep the presentation concrete and the benefits real; however, the applicability of the approach is general and we will touch on this in the discussion.

There are many advantages of *R*. We will focus on two. First, *R* is finely balanced to allow exploratory data analysis and interactive model development while also being a platform for statistical computing and data mining. As we will show, this is key for productivity and an element to set up (bit-perfect) reproducible models.

Second, it is comprehensive in the sense that most approaches to statistics and data mining are included in the tool or its contributed packages. Among other benefits, this allows you to easily run multiple model types on your data, ensuring compatibility with classic and often robust approaches while at the same time taking advantage of the latest developments and emerging industry standards.

Non-life insurance pricing is a well-known and well-established process and yet still a critical business issue. The standard for tariff analysis is generalised linear models. We first show how to develop such a model in *R*, including model selection and validation. We touch upon how to deploy the model (both scoring using the model and updating the model itself) while ensuring the results remain validated and reproducible.

Next we show how easy it is to extend the model to more complex techniques. In the interest of time we jump over intermediate approaches and go straight to ensemble models, which are possibly the state-of-the-art for high-performance models.

We are in no way advocating wholesale abandonment of classical approaches for modern techniques, ``black-box" or otherwise. Rather, we propose that you make use of both: continuity and understanding tempered with the results from the latest up-to-date methods. In the final part we cover some of these business issues to show how other insurers resolved them and what commercial benefits resulted. Examples include using the advanced models to restrict the validity domain of the classical approach (“risk we do not understand and will not insure”) and using them to create derived variables, such as interaction variables, to extend the domain of the GLM (“understanding complex risk”).

ReGenesees: symbolic computation for calibration and variance estimation

Diego Zardetto[1]

1. Istat – The Italian National Institute of Statistics
*Contact author: zardetto@istat.it

Keywords: Complex estimators, variance estimation, automated linearization, symbolic computation

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged *R* software for design-based and model-assisted analysis of complex sample surveys. It is the outcome of a long-term research and development project, aimed at defining a new standard for calibration, estimation and sampling errors assessment to be adopted in all Istat large-scale sample surveys. The system is distributed as Open Source Software, under the EUPL license. It can be freely downloaded from Istat website[1], as well as from JOINUP[2]. **ReGenesees** is rather different from existing estimation platforms developed by NSIs (mostly based on *SAS*) from both the application logic and the user experience standpoints. In a nutshell:

(1) User interaction with the new system takes place at a *very high level of abstraction*. **ReGenesees** users, indeed, no longer need to preprocess the survey data relying on ad-hoc programs; instead, they only have to feed the software with (i) the data as they are, plus (ii) *symbolic metadata* that describe the adopted sampling design and calibration model. At that point, it is up to the system itself to transform, in an automatic and transparent way, the survey data into the complex data structures required to solve the calibration problem and to compute estimates and errors.

(2) Besides Totals, and Absolute Frequency Distributions (linear estimators that are covered by all traditional platforms), **ReGenesees** allows to compute estimates and sampling errors with respect to Means, Ratios, Multiple Regression Coefficients, Quantiles, and, more generally, with respect to any *Complex Estimator*, provided it can be expressed as a differentiable function of Horvitz-Thompson or Calibration Estimators. It is worth stressing that such Complex Estimators can be defined in a completely free fashion: the user only needs to provide the system with the *symbolic expression* of the estimator as a mathematical function. **ReGenesees**, indeed, is able to automatically linearize such Complex Estimators, so that the estimation of their variance comes at no cost at all to the user.

Traditional estimation softwares did not give any support to the users in preparing auxiliary variables and population totals for calibration, nor in deriving the Taylor expansion of non-linear estimators and in computing the corresponding linearized variable for variance estimation purposes. As a consequence, ad-hoc (often very complex) programs for data preparation, transformation and validity check were developed and maintained outside the scope of the estimation system: a time-consuming and error-prone practice. **ReGenesees** frees its users from such needs, with an evident gain in terms of workload reduction, better usability and increased robustness against possible errors. Interestingly, both the innovative **ReGenesees** features sketched above leverage a peculiar strong point of the *R* programming language, that is its ability to process *symbolic information*.

---

[1] http://www.istat.it/it/strumenti/metodi-e-software/software/regenesees

[2] https://joinup.ec.europa.eu/software/regenesees/description

Big data exploration with tabplot

Martijn Tennekes[1*], Edwin de Jonge[1]

1. Statistics Netherlands
*Contact author: m.tennekes@cbs.nl

Keywords: Big data, visualization

The tableplot is an innovative visualization to explore large datasets (Tennekes et al., 2013). A tableplot is created by 1) sorting the data, 2) binning the data, 3) calculating mean values or category fractions, 4) visualizing it column-wise by a bar chart of mean values and a stacked bar chart of category fractions. The tableplot has been implemented in the *R* package **tabplot**. It includes a graphical user interface that uses the **shiny** package as well as the *javascript* **d3** package.

The **tabplot** package was introduced at the useR 2012 by a poster presentation. Recent developments include the increase of processing speed, which is in particular needed for the interactive interface, and the visualization of high cardinality categorical data. Furthermore, the **tabplot** package has been applied on two big data sources at Statistics Netherlands. The Dutch Virtual Census contains demographic information like age, gender, and household status about all 16,5 million Dutch inhabitants. The Dutch Labour and Benefits Database is a data source of all salaries and social benefits of the Dutch population, and contains about 100 million records on an annual basis.

The processing time consists of two major parts: sorting and aggregating. Both parts are implemented using the *R* package **ffbase** which uses *C* code. To increase the interactive performance, the sorting part is executed only once as a data preparation step. For each variable, the order of values is determined and stored. With these sorting orders, aggregation is fast, independent of the number of row bins. This makes it possible to explore datasets in a fast, interactive way. When even more speed is required, a uniform sample of the ordered data is drawn, aggregated, and plotted. Visualizing a sample is sufficient for initial data exploration. When a deeper look at the data is required, a larger sample and eventually the full dataset is processed.

The case studies at Statistics Netherlands emphasized the importance of visualizing categorical data. It is often worthwhile to categorize numeric variables, because this provides a better understanding of the data distributions per bin, as well as on the distribution of missing values along the bins. The core function `tableplot` offers flexibility to specify color palettes for categorical data. High cardinality categorical variables are visualized by grouping the categories, either by a specific aggregation scheme at hand or by merging neighboring categories uniformly so that the merged categories represent an equal number of original categories.

References

Tennekes, M., Jonge, E. de, Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots, *Journal of Data Science 11 (1)*, 43–58.

rwiot: An *R* package for Input-Output analysis on the World Input Output Database (WIOD)

Dong Guo[1*] **Valentin Todorov**[1]

1. United Nations Industrial Development Organization (UNIDO)
*Contact author: d.guo@unido.org

Keywords: input-output analysis, linkage, WIOD, structural change

International trade is becoming increasingly globalized and the products traded today are not produced in a single country but are the end-result of a series of steps carried out in many countries throughout the world. A basic method of quantitative economics which considers the macroeconomic activity as a system of interrelated goods and services is the Input-Output analysis [1]. It observes various economic sectors as a series of inputs (raw materials, supplies and services) and outputs (finished or semi-finished goods and services). In particular, it provides the tools to assess structural changes in the economy, in terms of linkages between economic sectors. The World Input-Output Database (WIOD) [3] is a new public data source which provides time-series of world input-output tables for the period from 1995 to 2009. National input-output tables of forty major countries in the world (covering about 90% of world GDP) are linked through international trade statistics.

Given the availability of this valuable statistical information, the analysts will need readily available and easy to use software for accessing and analyzing the data. The *R* package **rwiot** developed at UNIDO provides analytical tools for exploration of the various dimensions of the internationalization of production WIOD through time and across countries using input-output analysis. The package contains functions for basic (Leontief and Goshian inverse, backward and forward linkage, impact analysis) as well as advanced (vertical specialization) input-output analysis. Compositional data analysis techniques can be applied to study the interregional intermediate flows by sector and by region. The flexible *R* packaging mechanism is used to extend the functions with data, complete documentation and a large number of examples. In order to include the huge world input-output table (more than 250 MB) into the package a two-tear approach is used. The package **rwiot** includes the data for only one year (less than 15 MB) and thus can be downloaded from the CRAN repository and installed. After that a second package, **rwiotData** containing the complete table can be downloaded from an alternative web site. Special attention is given to the visualization of the analysis. The results can be presented in publication quality graphics as dot plots, radial plots or time series plots. The object oriented structure of the package allows for easy extension of the package functionality.

The presentation is illustrated with an example studying the manufacturing sectors and linking the world input-output tables to the UNIDO statistical database INDSTAT.

References

[1] Miller, R. and P. Blair (2009). *Input-Output Analysis: Foundations and Extensions*. Cambridge University Press.

[2] Nazara, S., D. Guo, G. J. Hewings, and C. Dridi (2004). Pyio: Input-output analysis with python. Technical report.

[3] Timmer, M., A. A. Erumban, R. Gouma, B. Los, U. Temurshoev, G. J. de Vries, I. Arto, V. A. A. Genty, F. Neuwahl, J. M. Rueda-Cantuche, A. V. J. Francois, O. Pindyuk, J. Poschl, R. Stehrer, and G. Streicher (2012). The world input-output database (WIOD): Contents, sources and methods. WIOD Background document.

Make Your Data Confidential with the sdcMicro and sdcMicroGUI packages

Alexander Kowarik[1], **Matthias Templ**[1,2,⋆]

1. Statistics Austria
2. Vienna University of Technology
⋆Matthias Templ: matthias@data-analysis.at

Keywords: Statistical disclosure methods, disclosure risk and data utility, S4 class package, graphical user interface

The demand of data from surveys, registers or other data sets containing sensible information on people or enterprises have increased significantly over the last years. However, before providing data to the public or to researchers, confidentiality has to be respected for any data set containing sensible individual information. Confidentiality can be achieved by applying statistical disclosure methods on the data.

We present a methodological and object-oriented approach to anonymize data - the **sdcMicro** package (Templ and Meindl, 2010; Templ et al., 2013), which includes all popular methods on statistical disclosure control. After specifying an S4-class *"sdcMicro"* object, all methods are directly applied to this object, whereas all necessary slots are filled in and updated automatically. For example, the disclosure risk and data utility estimates update whenever a method is applied on this object. This allows comparisons (how much a disclosure method influences the data utility/risk) and gives high usability into the hand of the user. Moreover, a reporting system is provided, whereas the whole process of anonymisation is summarized.

In addition, an highly interactive point and click graphical userinterface is implemented and available in the **sdcMicroGUI** package (Kowarik and Templ, 2013). All essential information is always updated and made visible to the user.

To allow for fast computations, all essential methods in the **sdcMicro** package are based on efficient `C++` code.

References

Kowarik, A. and M. Templ (2013). *sdcMicroGUI: Graphical user interface for package sdcMicro.* R package version 1.0.0.

Templ, M., A. Kowarik, and B. Meindl (2013). *sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package.* R package version 4.0.0.

Templ, M. and B. Meindl (2010). Practical applications in statistical disclosure control using R. In J. Nin and J. Herranz (Eds.), *Privacy and Anonymity in Information Management Systems*, Advanced Information and Knowledge Processing, pp. 31–62. Springer London. 10.1007/978-1-84996-238-4_3.

MRCV: A Package for Analyzing the Association Among Categorical Variables with Multiple Response Options

Natalie Koziol[1,*], **Christopher Bilder**[1]

1. Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE
*Contact author: nak371@neb.rr.com

Keywords: bootstrap, categorical data, correlated binary data, loglinear model, survey data

Multiple response categorical variables (MRCVs), also known as 'pick any' variables, summarize survey questions for which respondents are allowed to select more than one response option. For example, cancer survivors might select, from a predefined list, all of the treatments they have received over the past 5 years, and as a separate question, all of the side effects they have experienced during that same time period. Traditional methods for analyzing the association between categorical variables are not appropriate with MRCVs due to the within-subject dependence among responses. As such, alternative approaches have been proposed. Bilder and Loughin (2004) extended the Pearson chi-square statistic to test for marginal independence between two MRCVs. The authors describe three methods, including a nonparametric bootstrap approach, Rao-Scott second-order adjustment, and Bonferroni adjustment, that can be used in conjunction with the modified statistic to construct appropriate tests for independence. Bilder and Loughin (2007) introduced a more general loglinear modeling approach for analyzing MRCVs that relies on the use of marginal estimating equations. Bootstrap and Rao-Scott second-order adjustment methods are used to obtain appropriate standard errors and sampling distributions for model comparison statistics.

We developed the **MRCV** package (soon-to-be released to CRAN) to implement the methods proposed by Bilder and Loughin (2004, 2007). Our functions can be used to obtain parameter estimates and standard errors, perform model-comparison tests, calculate standardized residuals, and estimate model-predicted odds ratios. Other descriptive information is also available. We will use data from a survey of Kansas farmers to demonstrate our functions.

References

Bilder, C., and Loughin, T. (2007). Modeling association between two or more categorical variables that allow for multiple category choices. *Communications in Statistics—Theory and Methods, 36*, 433-451.

Bilder, C., and Loughin, T. (2004). Testing for marginal independence between two categorical variables with multiple responses. *Biometrics, 60*, 241-248.

Different tests on lmer objects (of the lme4 package): introducing the lmerTest package.

Alexandra Kuznetsova[1,*], Rune Haubo Bojesen Christensen[1], Per Bruun Brockhoff[1]

1. Department of Applied Mathematics and Computer Science. Technical University of Denmark Matematiktorvet Building 324, 2800 Kgs. Lyngby Denmark
*Contact author: alku@dtu.dk

Keywords: Satterthwaite, Kenward-Roger, degrees of freedom, linear mixed models, contrasts

One of the frequent wishes in *R*-sig-mixed-models is to get p-values for the `summary` and `anova` tables provided by **lme4** package (Bates et al., 2011). In **lmerTest** package (A. Kuznetsova et al., 2013) we have obtained `anova` and `summary` tables with p-values by implementing approximations to the degrees of freedom. We have implemented the Satterthwaite's approximation for denominator degrees of freedom as implemented in SAS/STAT® software `proc mixed` procedure (Fai et al., 1996) and also the type 3 hypothesis contrast matrix (J.H. Goodnight (1976), Goodnight, J. H. (1978),) in order to make tests for each fixed effect of the model thereby obtaining F statistics and corresponding p-values for each term of `anova` and `summary` tables. Furthermore type 3 `anova` with Kenward-Roger approximation for denominator degrees of freedom (based on the `KRmodcomp` function of the **pbkrtes**t package) can also be calculated which for more complex models and with a small sample size can give more reliable results.

The package also provides the new `step` function for lmer object. Starting with the most complex model it first eliminates non-significant random effects following by non-significant fixed effects (the elimination process of the fixed effects follows the principle of marginality) thereby finding the best by principle of parsimony model. For the final best model the function gives estimates for fixed and random effects as well as a post-hoc analysis: calculates least squares means with CI and differences of least squares means with CI. The `plot` function for the `step` function can visualize the post hoc analysis using barplots.

References

A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R-Version:1.1-0. http://cran.r-project.org/web/packages/lmerTest/index.html

Kuznetsova, A., Christensen, R.H.B., Bavay C. and Brockhoff, P.B. (2013). Automated Mixed ANOVA Modelling of sensory and consumer data. *To be submitted to: Food Quality and Preference.*

Bates, D., M. Maechler, and B. Bolker (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Fai, A. H. T. and Cornelius, P. L. (1996), Approximate F-Tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-Plot Experiments, Journal of Statistical Computation and Simulation, 54, 363–378.

J.H. Goodnight (1976). The General Linear Models. Proceedings of the First International SAS user's group, Cary, N.C.: SAS Institute Inc.

Goodnight, J. H. (1978), SAS Technical Report R-101, Tests of Hypotheses in Fixed-Effects Linear Models, Cary, NC: SAS Institute Inc.

Implementation of advanced polynomial chaos expansion in R for uncertainty quantification and sensitivity analysis

Miguel Munoz Zuniga [1*], **Jordan Ko**[2], **Yann Richet**[1]

1. Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-roses, France
2. Areva, Paris, France
*Contact author: mmunozzu@hotmail.com

Keywords: polynomial chaos expansion, sensitivity analysis, uncertainty propagation, sparse regression, quadrature

Polynomial chaos expansion is being increasingly used in the industry for uncertainty quantification applications. Initially, proposed by Weiner to represent random solution response with respect to Gaussian random variable inputs, polynomial chaos expansion (PCE) metamodel mimics the response of the solution over the random input parameter space. Its domain of application goes from quantile estimation, reliability analysis and solution optimization to sensitivity analysis and statistical moments quantification. Moreover, in the case of compuationally expensive models the PCE technique which maps random inputs and random outputs is a very useful and practical way of making computations tractable. One of its main appeal lies in its non-intrusive approach: the PCE metamodel can be constructed from samples considering the complete numerical model as a black-box. We implemented the non-intrusive PCE concept using three different methods for the expansion calculation: the Gauss quadrature, the adaptive sparse expansion and the adaptive sparse expansion based on least angle regression, gathered in one package entitled **GPC**. We will demonstrate the practical interest of the **GPC** R-package for the industrial, compare some common results obtained with the package **sensitivity** and present an application case using the new practical computer experiments scheme introduced in a related session.

dhglm & frailtyHL : R package for double hierarchical genralized linear models and frailty models

Maengseok Noh[1,⋆], Seungyoung Oh[2], Marek Molas[2], Youngjo Lee[2]

1. Department of Statistics, Pukyong National University, Busan, Korea
2. Department of Statistics, Seoul National University, Seoul, Korea
⋆Contact author: msnoh@pknu.ac.kr

Keywords: double hierarchical generalized linear models, frailty models, hierarchical likelihood, random effects models

Recently, some *R* packages have been developed for fitting hierarchical generalized linear models (HGLMs) of Lee and Nelder (1996) by using h-likelihood procedures such as **hglm** package and **HGLMfit** package. HGLMs were developed from a synthesis of generalized linear models, random-effect models and structured dispersion models. Double hierarchical generalized linear models (DHGLMs) are futher extension of HGLMs by allowing random effects in various components in HGLMs like randome effects in the despersion model. The **dhglm** package can be used to fit DHGLMs.

Frailty models are extension of proportional hazard model (Cox, 1972) to consider correlations among data. They have been widely used in analysis of correlated survival data, for example, survival time measured repeatedly on the same indivisual. Previously, some researcher have developed several *R* functions such as the *coxph()* function in the **survival** package, the *coxme()* function in the **coxme** package, the *phmm()* function in the **phmm** package and the *frailtyPenal()* function in the **frailtypack** package. The **frailtyHL** package implements h-likelihood estimation procedures for fitting frailty models.

This presentation will introduce **dhglm** package and **frailtyHL** package, show how to work these packages through basic usage, and finally illustrate their power with a few advanced examples.

rknn: an R Package for Parallel Random KNN Classification with Variable Selection

E. James Harner[1,⋆], **Shengqiao Li**[2], **Donald A. Adjeroh**[3]

1. Department of Statistics, West Virginia University
2. UPMC Health Plan
3. Lane Department of Computer Science and Electrical Engineering, West Virginia University

⋆Contact author: jharner@stat.wvu.edu

Keywords: Machine Learning, K-Nearest Neighbor, High Dimensional Data, Parallel Computing

Random KNN (RKNN) is a novel generalization of traditional nearest-neighbor modeling. Random KNN consists of an ensemble of base k-nearest neighbor models, each constructed from a random subset of the input variables. A collection of *r* such base classifiers is combined to build the final Random KNN classifier. Since the base classifiers can be computed independently of one another, the overall computation is *embarrassingly parallel*.

Random KNN can be used to select important features using the RKNN-FS algorithm. RKNN-FS is an innovative feature selection procedure for "small n, large p problems." Empirical results on microarray data sets with thousands of variables and relatively few samples show that RKNN-FS is an effective feature selection approach for high-dimensional data. RKNN is similar to Random Forests (RF) in terms of classification accuracy without feature selection. However, RKNN provides much better classification accuracy than RF when each method incorporates a feature-selection step. RKNN is significantly more stable and robust than Random Forests for feature selection when the input data are noisy and/or unbalanced. Further, RKNN-FS is much faster than the Random Forests feature selection method (RF-FS), especially for large scale problems involving thousands of variables and/or multiple classes.

Random KNN and feature selection algorithms are implemented in an *R* package **rknn**. The time complexity of the algorithm, including feature selection, is $O(rkpn \log n)$, assuming the number of variables randomly selected in a base classifier is $m = \log p$. This choice of m, in contrast to $\sqrt{p}$, reduces the time complexity from exponential time to linear time. However, it is important to choose r sufficiently large to ensure adequate variable coverage. By paralleling the code in **rknn**, the time can be reduced linearly depending on the number of cores or compute nodes. The basic **rknn** package has been extended to support parallel processing using the **parallel** package. The code detects whether the system is Posix-based and then determines whether a "FORK" or "PSOCK" cluster is formed. Parallelization is also supported using `mclapply`. We will show how to apply the Random KNN method via the parallelized **rknn** package to high-dimensional genomic data.

References

Li S, Harner EJ, Adjeroh DA (2011). Random KNN feature selection—a fast and stable alternative to Random Forests. *BMC Bioinformatics*, 12(1):450.

Patterns of Multimorbidity: Graphical Models and Statistical Learning

Matthias Eckardt[1,2,⋆]

1. University Medical Center Hamburg-Eppendorf, for MultiCare Study Group
2. Hamburg Center for Health Economics
⋆Contact author: m.eckardt@uke.de

Keywords: Graphical Models, Regression Trees, Finite Mixtures, Statistical Learning

Multimorbidity is defined as the coexistence of three or more chronic diseases per individual. Especially among older people aged above 65 years over 60% of the population are suffering from coexisting single diseases like Parkinson, diabetes melitus, asthma, or joint arthrosis. Thus, individuals with multiple chronic conditions consume a disproportionally large share of total health services and healthcare expenditures. As a result of demographic change, the prevalence of multimorbidity is expected to substantially increase in future decades.
Based on survey data of n = 1050 patients suffering from arbitrary multiple disease combinations we aim to detect the most influential multimorbidity patterns affecting healthcare costs. Taking 42 single diseases and certain sociodemographic variables into account we applied graphical models to visualize the structure and conditional dependencies of the data. Besides this, conditional inference trees, Random Forests, LASSO and finite mixtures of generalized linear models were used. As a result, certain subgroups and most influential diseases have been found. Analysis was based on the *R* packages **party, bnlearn, glmnet** and **flexmix**.

References

Bishop, Christopher M. (2007). Pattern Recognition and Machine Learning. Springer.

Hastie, Trevor and Tibshirani, Robert and Friedman, J. H. (2001).The elements of statistical learning: data mining, inference, and prediction. Springer.

Lauritzen, Steffen (1996). Graphical Models. Oxford.

McLachlan, Geoffrey and Peel, David (2000). Finite Mixture Models. John Wiley & Sons.

Pearl, Judea (2009). Causality: Models, Reasoning and Inference. Cambridge.

ExactSampling: risk evaluation using exact resampling methods for the k Nearest Neighbor algorithm

Kai Li[1], **Alain Célisse**[2,4], **Michel Koskas**[1], **Tristan Mary-Huard**[1,3,⋆]

1. AgroParisTech/INRA, UMR 518 MIA, F-75005 Paris, France.
2. Laboratoire de Mathématique Paul Painlevé, UMR 8524 CNRS - Université Lille 1, France.
3. UMR de Genetique Vegetale, INRA, Universite Paris-Sud, CNRS, Gif-sur-Yvette, France
4. Equipe-projet MODAL, INRIA Lille, France

⋆Contact author: maryhuar@agroparistech.fr

Keywords: Machine learning, cross-validation, bootstrap

The k-Nearest Neighbor algorithm (kNN) is a popular method to perform either regression or classification. It consists in predicting the response of a new observation according to the k closest observations in the training sample. While simple, the kNN algorithm demonstrated good performances on many applications (Hastie et al., 2001), and has already been implemented in several *R* packages (**class**,**FNN**).

The performance of the kNN algorithm highly depends on the tuning of parameter k, that should be performed adaptively to the data at hand. To do so, resampling strategies such as Bootstrap or Leave-p-out (LpO) cross-validation can be used to estimate the prediction performance obtained with different values of k, and select the optimal value k^* that minimizes the prediction error rate. However, the computational cost of such strategies is prohibitive. In practice one often needs to limit the number of resamplings as the training sample size gets large, yielding poor approximation of the actual risk. Recently computational shortcuts have been derived to compute the exact LpO or Bootstrap risk estimators in a short computational time, for instance linear with respect to the number of observations in the case of LpO, whatever p (Celisse, 2011). These efficient strategies are now implemented in a package called **ExactSampling**. In this package additional computational shortcuts are also provided in settings where exacts formulas are not available for the bootstrap estimator. In particular the approximation level can be specified beforehand by the practitioner.

From an algorithmic point of view, we improve the computational time of the nearest neighbor search step, which is the most consuming one. This is done thanks to a new combination of classical algorithms. All functions have been developed using the *C* programming language. From a theoretical point of view, the package may be used to investigate the properties of resampling methods that are still poorly understood. In practice **ExactSampling** can be of great help to tune the k parameter of the kNN when dealing with large dataset: while a naive implementation of the leave-10-out cross validation procedure would be totally inefficient to deal with a dataset of size 100 (hundreds of hours of computational time), the `knn.cv` function of the **ExactSampling** package computes this estimator within a few minutes for datasets of size 100,000.

References

Celisse, A. Mary-Huard, T. (2011). Exact cross-validation for knn and applications to passive and active learning in classification. *JSFdS 152*(3).

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

Classifying High-Dimensional Data with the The HiDimDA package

A. Pedro Duarte Silva[1,2*]

1. Faculdade de Economia e Gestão / Catholic University of Portugal
2. Centro de Estudos em Economia e Gestão
*Contact author: psilva@porto.ucp.pt

Keywords: Discriminant Analysis, High Dimensionality, Variable Selection, Large Covariance Estimation.

Classical methods of supervised classification often assume the existence of a training data set with more observations than variables. However, nowadays many applications work with data bases where the total number of original features is much larger than the number of available data units. Nevertheless, in most high-dimensional classification problems the majority of the original variables do not contribute to distinguish the underlying classes, but the number of useful features is often still comparable to, or even larger than, the number of available training sample observations.

Therefore, effective classification methodologies for these applications require scalable methodologies of variable selection, and classification rules that work well in the few observations / many variables settings. A common strategy to achieve the later goal is to adopt rules that ignore the dependence structure of the data (*e.g,*. Fan and Fan (2008)). Recent proposals (*e.g.*, Thomaz, Kitani, and Gilies (2005), Fisher and Sun (2011), Duarte Silva (2011)) rely instead on rules based on estimators of covariance matrices with good statistical properties under such conditions.

In this presentation, I will describe the **HiDimDA** (*High Dimensional Discriminant Analysis*) *R* package, available on CRAN, that implements several routines and utilities for supervised *k*-group classification in high-dimensional settings. **HiDimDA** includes routines for the construction of classification rules with the above mentioned properties, methods for predicting new observations, as well as cross-validation and variable selection utilities.

HiDimDA can be used to construct, apply and assess k-group classification rules for problems with several thousand variables, dealing effectively with the problems of high dimensionality, and including rules that do not ignore the dependence structure of the data.

References

Duarte Silva, A.P. (2011) "Two-Group classification with High-Dimensional correlated data: A factor model approach." *Computational Statistics and Data Analysis* **55** (11), 2975-2990.

Fan, J. and Fan, Y. (2008) "High dimensional classification using Features Annealed Independence Rules." *The Annals of Statistics* **38**. 2605-2637.

Fisher, T.J. and Sun, X. (2011) "Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix." *Computational Statistics and Data Analysis* **55** (5), 1909-1918.

Thomaz, C.E; Kitani, E.C. and Gilies, D.F. (2006) "A maximum uncertainty LDA-based approach for limited sample size problems – with application to face recognition." *Journal of the Brazilian Computer Society* ***12*** *(2), 7-18.*

Groupon Impact Report: Using *R* To Power Large-Scale Business Analytics

Raju Balakrishnan[1], **Natalia Corominas**[1], **Latife Genc-Kaya**[1], **Amit Koren**[2], **Kamson Lai**[1,⋆], **Francisco Larrain**[1], **Gaston L'Huillier**[1], **Cristian Orellana**[1]

1. Groupon, 3101 Park Blvd, Palo Alto, California, USA
2. Groupon, 600 W. Chicago Ave, Chicago, Illinois, USA
⋆Contact author: kamson@groupon.com

Keywords: Business Analytics, Big Data, Local Businesses

The Groupon Impact Report is a web service designed to provide Groupon's merchant partners with in-depth analytics pertaining to the performance of their Groupon campaigns. A suite of metrics presented in an easy-to-understand website helps merchant partners gain insights into many aspects of their businesses, including customer acquisition, customer retention, and return on investment (ROI). Launched in January 2013, the Impact Report currently serves over 80,000 of Groupon's North American merchant partners.

The engine performing the data analysis for the Impact Report is implemented in *R*. We chose to develop in *R* because of 1) the ease and speed of development for data scientists, 2) ability to interface with SQL databases and mash-ups with different data sources, and most importantly 3) excellent support of statistical and machine learning tools. The production *R*-engine queries a database built from a variety of sources for the necessary data. After computing the metrics, the results are written to a database dedicated to serving the metrics via a REST API used by internal clients. The *R*-engine processes a large amount of data, computing and updating nearly 24 million metrics on a daily basis. Since metric computation for each merchant is separate and independent, parallelization can be achieved by running multiple concurrent calculations. This makes the system easily scalable.

The foundation of the Impact Report is Groupon's merchant data. Purchase and coupon redemption data provide insights into customer demographics and acquisition. Post-redemption surveys measure customer satisfaction and loyalty. In-store transaction data track customer behavior and provide an estimate of the overall financial impact. Leveraging this powerful data set, the Impact Report presents merchants with a comprehensive view of their Groupon campaigns and beyond. Another major component of the Impact Report is a model that predicts metrics for merchants with missing or incomplete input data. Regression models are trained for a number of metrics using as input historical data and merchant features such as business category and merchant quality. This predictive model makes it possible to expand coverage of the Impact Report to nearly 100% of Groupon's North American merchant partners.

The future of the Impact Report lies in real-time applications. It is desirable to update metrics as soon as new data arrive. Another interesting application is to allow users to interact with the data, for example, by adjusting inputs specific to their businesses. Implementing *R* in the backend of a real-time web application will present unique challenges, especially in the areas of performance and web integration.

A demo of the Impact Report is available at https://merchants.groupon.com/demo#/deals/impact.

Statistics with Big Data: Beyond the Hype

Joseph Rickert[1,*]

1. Revolution Analytics
*Contact author: joseph.rickert@revolutionanalytics.com

Keywords: R, statistics, big data, Hadoop

This past year the hype surrounding “Big Data” has dominated the “Data Science” world and made quite a stir in the main-stream media. Entire industries have grown up around technologies such as Hadoop and MapReduce promising astonishing insights from “crunching” large amounts of data. Putting hyperbole aside, what are the theoretical and practical challenges involved in working with very large data sets and what tools exist in R to help meet these challenges? In this talk, I will offer some ideas about how to think of big data from a statistical point of view, make some suggestions on computer architectures for facilitating working with *R* and large data sets, and show some examples of R code used to analyze large data sets including **biglm**, **Rhadoop** and **RevoScaleR code**. I will also illustrate how very large data sets are forcing developers to rethink basic algorithms by describing the rxDTrees algorithm in the **RevoScaleR** package that builds classification and regression trees on histogram summaries of the data.

Using survival analysis for marketing attribution (with a big data case study)

Andrie de Vries[1,2,⋆]

1. Director of Business Service - Europe, Revolution Analytics
2. Author of R for Dummies (de Vries and Meys (2012))

⋆Contact author: andrie@revolutionanalytics.com

Keywords: Survival analysis, Attribution modelling, Big data

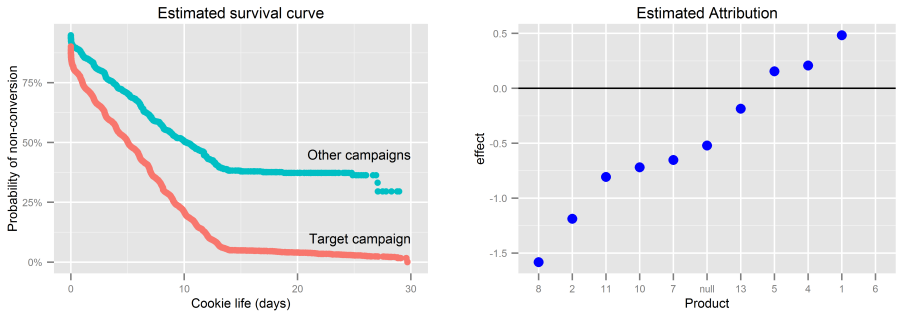
A central question in advertising is how to measure the effectiveness of different ad campaigns. In online advertising, including social media, it is possible to create thousands of different variations on an ad, and serve millions of impressions to targeted audiences each day.

Rather too often, digital advertisers use the **last click attribution model** to evaluate the success of campaigns. In other words, when a user clicks on an ad impression, only the very last event is deemed as signficant. This is convenient but doesn't help in making good marketing decisions.

Survival analysis is widely used in the modeling of living organisms and time to failure of components, but Chandler-Pepelnjak (2010) proposed to use survival analysis for marketing attribution analysis. The insight is that each event in a clickstream (the series of impressions and clicks) gives an indication of whether the clickstream is still alive. Only at the final click (or impression), when the user converts to purchase, does the clickstream "die". These effects can be estimated with the Cox proportional hazards model, using the function `coxph()` in package **survival**. This allows the estimation of time effects, the effects of clicks, media format and other factors.

We illustrate this approach using a small dataset of impression and click data of 750K events, representing 135K purchase events, mainly arising from advertising on Facebook and other social media.

In conclusion, we provide a big data case study, showing how Upstream used Revolution Analytics to process 50 million survival models each day.



References

Chandler-Pepelnjak, J. (2010, May). *Modeling Conversions in Online Advertising*. Ph. D. thesis, The University of Montana.

de Vries, A. and J. Meys (2012). *R for Dummies*. John Wiley and Sons.

Big Data Analytics – Scaling *R* to Enterprise Data

Mark Hornick

Oracle
mark.hornick@oracle.com

Keywords: big data, scalability, Oracle R Enterprise, advanced analytics, Oracle

Enterprise data commonly reside in relational databases and increasingly in the Hadoop Distributed File System. It is no longer competitive to rely solely on traditional data warehouse data for business insights. Enterprises now need to analyze large volumes of social media data, sensor data, blog and web data in combination with enterprise data warehouse data. For the new generation of data analysts, leveraging *R* in this new space can be a real productivity gain. However, this gain is only realized when the underlying technologies provide scalability and performance, combined with streamlined application development and production deployment.

In this session, we focus on leveraging the power of *R* with data stored in Oracle Database and Hadoop, while leveraging both as powerful compute engines supporting Big Data analytics. We step through the phases of a big data analytics project, starting with interactive analysis, reusable *R* script creation and management, and finally production deployment. At each point, we explain the capabilities of Oracle R Enterprise and Oracle R Connector for Hadoop that enable that solution.

References

Oracle (2013). Oracle R Enterprise, http://www.oracle.com/technetwork/database/options/advanced-analytics/r-enterprise.

Oracle (2013). Oracle R Connector for Hadoop, http://www.oracle.com/us/products/database/big-data-connectors.

Using R for exploring sampling designs at Statistics Norway

Susie Jentoft[1*], **Johan Heldal**[1]

1. Division for Methods, Statistics Norway
* Contact author: Susie.Jentoft@ssb.no

Keywords: sampling, official statistics, Rcmdr, RcmdrPlugin.sampling

The use of *R* at Statistics Norway is mostly restricted to the Division for Methods. It is used as an analytic, analysis and development tool. However, the benefits of *R* over other software packages are growing and we are looking for more ways to integrate the software into our everyday processes.

One area of interest for integration is in the planning phase of survey sampling. We often use complex sampling designs, particularly in face-to-face interview surveys, to cluster the participants in manageable and cost-effective groups. However, clustering participants generally reduces the precision of the estimates from the survey. Investigating this balance between cost and accuracy is an integral part of the planning process. We have been developing a package called **RcmdrPlugin.sampling** as a tool for exploring sampling designs and selecting samples. This package uses tools from the **sampling** package and builds them into the **Rcmdr** interface. This is to give the package a user friendly interface that can appeal to people with little experience of *R*. It provides a platform for selecting a simple random sample, stratified samples and multi-stage samples with or without stratification. If key variable and cost data are available, estimates for variances and survey costs can be calculated for different stratification and clustering designs.

This package has been designed as a versatile and useful tool for both Statistics Norway and others who deal with planning sample surveys. We hope over time to continue to add to the functionality of the package and to add extra features for enhanced visualisation of the proposed sample.

Application of R in Crime Data Analysis

Anna Dyga[1*], Monika Sławińska[1**]

1.Kielce University of Technology
*Contact author: anna_dygaa@o2.pl
** Contact author: monikaslawinska@o2.pl

Keywords: Crime, Forecasting,Decision Tree,

The use of *R* in context of crime data analysis certainly contrasts with its most popular application – in finance, economics, or planning and prediction which are largely connected to trade or services. It is common knowledge that in many countries crime rate has been on the rise over the last few years and it is a complicated issue that governments struggle to solve. The aim of our presentation is to introduce the application of *R* program in the police's statistics and show its numerous advantages. *R* is an open source program and the use of it is fairly simple and does not necessarily require advanced IT abilities. This makes it an ideal means to deal with the enormous amount of data gathered by the police, to simplify its processing and interpretation. For the sake of this presentation, we have analyzed two data sets with the use of *R*. The first set contains data about crimes committed in 2010 in particular countries. The second set is concerned with data about crimes in all the regions of Poland. To analyze the data, we have used easily-comprehensible `histograms` which show the crime rate in every country. Besides, to depict which crimes are prevalent and which are committed less often, we have used `mosaic plots` available in **rattle** package. However, the main goal of our presentation is to specify the most dangerous areas and to this end we have employed `decision trees` with the help of packages **tree** and **rpart**. We have used this particular tool because it is lucid and greatly simplifies the reasoning. As countries included in the first data set are at various stages of development (developed, developing and least-developed), it have been possible to try to point out a relationship between the economic advancement and citizens' tendency to commit crimes. What is more, we have discovered whether types of crimes committed in those least-developed countries are in any way similar to or different from those which dominate in countries with high-functioning economy. For greater clarity, presentation of our findings and interpretation of them is followed by maps depicting crime rates in particular areas both in Poland and around the world. Moreover, the analysis of the second data set has enabled us to establish the extent of crime in Poland in comparison to other countries, and to learn whether Poland is among the countries of higher crime rate or it is relatively safe to visit the country. It is necessary to point out that *R* program can be further employed in the analysis of data concerning smaller units, namely city districts – a beneficial method especially in case of metropolises with diversified national and cultural makeup, such as London or New York City where multiculturalism may be a source of tension between people and create conditions conductive to crime rise. We are convinced that application of *R* would make crime forecasting easier and increase the ability to deal with crime on local and global scale.

References

MSWiA's Department of Analysis and Supervision. (2011) Report on Security Situation in Poland in 2010 (original in Polish).
Kabacoff, I. Robert. (2012) http://www.statmethods.net/advstats/cart.html.
Kopaczewska, Katarzyna. (2009) Econometric and Spatial Statistics with the Use of R CRAN.(original in Polish).
Yanchang Zhao(2012). R Data Mining: Examples and Case Studies. http://www.rdatamining.com/
http://www.nationmaster.com.

Maps can be rubbish for visualising global data : a look at other options.

Andy South

Consultant, Norwich, UK. southandy@gmail.com

Keywords: maps, global data, visualisation, rworldmap, R

Global maps can be beautiful and people (myself included) love to see them and put them in their academic papers. I've even developed an *R* package **rworldmap** focussed on making global maps (South, 2011). A map, however, is not necessarily the best way of communicating global data. Here I will explore potential disadvantages to using maps to communicate global data and outline solutions and alternatives, showing how to implement these in **rworldmap** and other *R* packages.

I will concentrate on global country-referenced data which has become increasingly available over the past ten years as a result of improved remote sensing, better storage and reporting of national statistics and an increasing fondness for the creation of composite indices such as the UN Human Development Index. Hans Rosling has accelerated this process for over 5 years with his beautiful and inspiring gapminder software, which itself demonstrates the power of taking world data away from the map. Unlike gapminder I will focus on static graphics for academic papers and the print media, but I will show briefly how the excellent **shiny** and **googleVis** packages can be used to create interactive visualisations of country referenced data.

The great resources in *R* for accessing such global country referenced data will be demonstrated briefly. For example the packages **FAOSTAT** and **WDI** make it easy to access country-referenced data from the FAO and World Bank Development Indicators respectively.

Data visualisation principally uses symbols, size, positioning and colour to communicate quantitative and qualitative information. World maps instantly communicate information about global patterns but aspects of the data can be lost due to differences in the size of countries and perceptual difficulties in comparing the size and/or colour of symbols not placed on a regular xy axes. I will demonstrate how *R* can be used to address these issues.

This talk will complement the mapping tutorial I am giving at the start of useR 2013. Attendees at both will get different information, however attendance at either is not necessary for the other.

References

Rosling, H. et al. (2005-2012). Gapminder : for a fact-based worldview. http://www.gapminder.org/.

South, A.B. (2011) rworldmap: A New R package for Mapping Global Data. The R Journal 3, 35-43. http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf

The use of demography package for population forecasting

Han Lin Shang[1,⋆]

1. University of Southampton
⋆Contact author: H.Shang@soton.ac.uk

Abstract

This paper describes some functional data models that are currently available in the **demography** package, for forecasting mortality, fertility and net migration. The forecasts of these demographic components allow us to obtain forecasts of population through a cohort-component projection model. The population forecasts can be of great importance to demographers, social statisticians, and policy makers. The implementation of these methods is illustrated by using the historical data of United Kingdom from 1975 to 2009.

Keywords: functional data analysis, functional principal component analysis, nonparametric smoothing, population forecasting

Shape constrained additive modelling in R

Natalya Pya

University of Bath, Department of Math Sciences, Bath BA2 7AY, UK
E-mail: n.y.pya@bath.ac.uk

Keywords: Monotonic smoothing, Convex smoothing, Generalized additive model

An R package **scam** is presented for generalized additive modelling under shape constraints on the component functions of the linear predictor of the GAM. Models can contain multiple shape constrained and unconstrained terms as well as shape constrained bivariate smooths. The shape constraints considered include a variety of monotonic functions such as increasing/decreasing, convex/concave, increasing/decreasing and convex, increasing/decreasing and concave. Also bivariate functions with monotonicity constraints along both covariates or only along only a single direction are considered. The shape constrained terms are represented by mildly non-linear extensions of P-splines. The model set up is the same as in the R-library **mgcv(gam)** (Wood, 2006) with the added shape constrained smooths, so the unconstrained smooths can be of more than one variable, and other user defined smooths can be included. `scam` and `plot.scam` functions are based on the `mgcv(gam)` and `mgcv(plot.gam)` and similar in use. `summary.scam` allows to extract the results of the model fitting in the same way as in `summary.gam`. Penalized likelihood maximization based on Newton-Raphson method is used to fit the model together with the automatic multiple smoothing parameter selection by GCV or UBRE/AIC.

References

Wood, S. (2007). *Generalized Additive Models. An Introduction with R*. Chapman & Hall.

Semiparametric bivariate probit models in *R*: the SemiParBIVProbit package

Rosalba Radice[1,⋆], **Giampiero Marra**[2]

1. Department of Economics, Mathematics and Statistics, Birkbeck, Univerity of London, WC1E 7HX London, United Kingdom
2. Department of Statistical Science, Univerity College London, WC1E 6BT London, United Kingdom
⋆Contact author: r.radice@bbk.ac.uk

Keywords: Bivariate probit models, penalized regression spline, *R*.

Bivariate probit models are a natural extension of the probit model where two binary equations are allowed to be dependent (Greene, 2012). The strength of this class of models lies in its ability to obtain efficient and consistent estimates of the parameters of interest in the presence of unobserved confounders (i.e. variables that are associated with both covariates and response). Traditional bivariate probit models assume pre-specified covariate-response relationships hence giving rise to a potential problem of residual confounding due to unmodelled non-linearities.

Marra and Radice (2011) introduced a model fitting procedure which allows us to estimate the parameters of bivariate probit models that include smooth functions of continuous covariates. The algorithm is based on the penalized maximum likelihood framework. In this talk, we will discuss this framework as well as some new extensions which allow us to estimate such models in the presence of non-random selection and test the hypothesis of absence of unobserved confounders. The methods are implemented in the *R* package **SemiParBIVProbit** (Marra and Radice, 2013), which will be demonstrated using fictitious and real data.

References

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.

Marra, G. and R. Radice (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics 39*, 259–279.

Marra, G. and R. Radice (2013). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.2-6.

"RobExtremes": Robust Extreme Value Statistics — a New Member in the RobASt-Family of *R* Packages

Peter Ruckdeschel[1,2,⋆], **Matthias Kohl**[3], **Nataliya Horbenko**[4]

1. Fraunhofer ITWM, Kaiserslautern, Germany
2. Kaiserslautern University, Germany
3. Furtwangen University, Germany
4. KPMG

⋆Contact author: peter.ruckdeschel@itwm.fraunhofer.de

Keywords: robustness, extreme value statistics, diagnostics, operational risk, hospital length of stay

As a software offspring of VW-foundation funded project *"Robust Risk Estimation"*, we present *R* package **"RobExtremes"**, covering the scope of our project and building up on the framework of the `distrXXX` and `RobAStXXX` families of packages implemented and maintained—together with varying coauthors—by the present authors.

Starting with package **"distr"**, providing an object-oriented framework for probability distributions, and extended by packages **"distrEx"** and **"RandVar"** with functionals and random variables acting on these distribution objects, we have set up an arithmetics for probability models. In particular, in package **"distrMod"**, we have built up an infrastructure for general smooth parametric models such that you can write, e.g. `MLEstimator(data, model)`.

In a further step, this approach has then been extended to cover the setup of infinitesimally robust statistics as presented in detail in Kohl et al. (2010). The corner stones are packages **"RobAStBase"** and **"ROptEst"**, the former including general concepts of robust statistics such as influence curves and corresponding diagnostics, the latter general infrastructure for optimally-robust estimators, respectively.

In the `RobAStXXX` family of packages, we have implemented optimally robust estimation in the infinitesimal setup of Rieder (1994), i.e., L_2-differentiable parametric models, shrinking neighborhoods, etc. By our general approach we may employ **one** algorithm for a large class of probability models, thus avoiding redundancy and simplifying maintenance.

Package **"RobExtremes"** implements the general LD estimators introduced in Marazzi and Ruffieux (1999), in particular including the high-breakdown point estimators medSn, medQn, and medkMAD discussed in Ruckdeschel and Horbenko (2012). In addition, an interpolation technique is applied to speed-up computation of the optimally-robust estimators MBRE, OMSE, RMXE.

We demonstrate this, together with corresponding diagnostics at some real data sets from the context of hospital length of stay and operational risk of a bank, as considered within our project.

References

Horbenko, N., P. Ruckdeschel, and T. Bae (2011). Robust Estimation of Operational Risk. *The Journal of Operational Risk 6*(2), 3–30.

Kohl, M. and P. Ruckdeschel (2010). R package distrMod: Object-Oriented Implementation of Probability Models. *Journal of Statistical Software 35*(10), 1–27.

Kohl, M., P. Ruckdeschel, and H. Rieder (2010). Infinitesimally Robust Estimation in General Smoothly Parametrized Models. *Statistical Methods and Applications 19*, 333–354.

Marazzi, A. and C. Ruffieux (1999). The truncated mean of asymmetric distribution. *Computational Statistics & Data Analysis* (32), 79–100.

Rieder, H. (1994). *Robust asymptotic statistics*. Springer Series in Statistics. Springer.

Ruckdeschel, P. and N. Horbenko (2011). Optimally-Robust Estimators in Generalized Pareto Models. *Statistics*. doi: 10.1080/02331888.2011.628022.

Ruckdeschel, P. and N. Horbenko (2012). Yet another breakdown point: EFSBP—illustrated at shape-scale models. *Metrika 8*(75), 1025–1047.

Generalized Bradley-Terry Modelling of Football Results

Heather Turner[1,2], **David Firth**[1] **and Greg Robertson**[1]

1. University of Warwick, UK
2. Independent statistical/R consultant

*Contact author: ht@heatherturner.net

Keywords: Statistical modelling, sport

For a regular sporting contest, such as a football league, records of match outcomes from the current and previous seasons provide a wealth of data for statistical inference. The Bradley-Terry model (Bradley and Terry, 1952) provides an intuitive way to analyse such data, in which the odds of team i beating team j are modelled by the ratio of the respective abilities, $\alpha_i, \alpha_j > 0$:

$$\text{odds}(i \text{ beats } j) = \frac{\text{pr}(i \text{ beats } j)}{\text{pr}(j \text{ beats } i)} = \frac{\alpha_i/(\alpha_i + \alpha_j)}{\alpha_j/(\alpha_i + \alpha_j)} = \frac{\alpha_i}{\alpha_j}.$$

A disadvantage of the Bradley-Terry model is that it assumes that the outcome is binary, that is, it does not allow for the possibility of a draw. Draws are a common result in football matches and therefore it is important to gain whatever information we can from such results. One possibility is to use the extension of the Bradley-Terry model proposed by Davidson (Davidson, 1970), in which the probability of a tie is incorporated as follows:

$$\begin{aligned} \text{pr}(\text{tie}) &= \frac{\nu\sqrt{\alpha_i\alpha_j}}{\alpha_i + \alpha_j + \nu\sqrt{\alpha_i\alpha_j}} \\ \text{pr}(i \text{ beats } j \mid \text{not tied}) &= \frac{\alpha_i}{\alpha_i + \alpha_j} \end{aligned}$$

Like the standard Bradley-Terry model, the Davidson model can be formulated as a log-linear model and hence is straight-forward to fit via `glm`. However, the Davidson model, with its single parameter for modelling the probability of a draw, severely restricts the relationship between draw probabilities and the estimated relative abilities of teams.

In this talk we propose a generalization of the Davidson model, which affords greater flexibility in modelling the probability of a draw. The model is formulated as a generalized nonlinear model, which we implement in R using the **gnm** package (Turner and Firth, 2012). We illustrate particular cases of the extended model, using results of football matches from the English first division.

References

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika 39*, 324–45.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Amer. Statist. Assoc 65*, 317–328.

Turner, H. and D. Firth (2012). *Generalized nonlinear models in R: An overview of the gnm package*. R package version 1.0-6.

Copula sample selection modelling using the *R* package SemiParSampleSel

Giampiero Marra[1,*], **Rosalba Radice**[2], **Małgorzata Wojtyś**[1,3]

1. Department of Statistical Science, Univerity College London, WC1E 6BT London, United Kingdom
2. Department of Economics, Mathematics and Statistics, Birkbeck, Univerity of London, WC1E 7HX London, United Kingdom
3. Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warszawa

[*]Contact author: giampiero.marra@ucl.ac.uk

Keywords: Copula, non-random sample selection, penalized regression spline, selection bias, *R*.

Sample selection models deal with the situation in which the observations available for statistical analysis are not from a random sample of the population. This occurs when individuals have selected themselves into (or out of) the sample based on a combination of observed and unobserved characteristics. Estimates based on models that ignore such a non-random selection may be biased. The estimation of such models is based on a binary equation, which describes the selection process, and an outcome equation, which is used to examine the substantive question of interest. Classic sample selection models assume a priori that continuous covariates have a pre-specified linear or non-linear relationship to the outcome, and that the distribution linking the error terms of the two equations is bivariate normal.

The literature on sample selection modeling is vast and many variants have been proposed (see, for instance, Marra and Radice (2013) and references therein). For example, the *R* packages **sampleSelection** (Toomet and Henningsen, 2012) and **bayesSampleSelection**, which is available at `http://www.unigoettingen.de/en/96061.html`, implement several sample selection models. However, both packages make the assumption of bivariate normality between the model equations and **sampleSelection** assumes a priori that continuous regressors have pre-specified linear or non-linear relationships to the responses. We introduce the *R* package **SemiParSampleSel** (Marra et al., 2013) to deal simultaneously with non-random sample selection, non-linear covariate effects and non-normal bivariate distributions between the model equations. The core algorithm is based on the penalized maximum likelihood framework proposed by Marra and Radice (2013) for the bivariate normal case.

References

Marra, G. and R. Radice (2013). Estimation of a regression spline sample selection model. *Computational Statistics and Data Analysis*.

Marra, G., R. Radice, and M. Wojtys (2013). *Semiparametric Sample Selection Modelling with Continuous Response*. R package version 1.0.

Toomet, O. and A. Henningsen (2012). *sampleSelection: Sample Selection Models*. R package version 0.7-2.

Robust model selection for high-dimensional data with the *R* package robustHD

Andreas Alfons*

Erasmus School of Economics, Erasmus University Rotterdam
*Contact author: alfons@ese.eur.nl

Keywords: Outliers, Robust least angle regression, Sparse least trimmed squares, Variable selection, *C++*

In regression analysis with high-dimensional data, variable selection is an important step to (i) overcome computational problems, (ii) improve prediction performance by variance reduction, and (iii) increase interpretability of the resulting models due to the smaller number of variables. However, robust methods are necessary to prevent outlying data points from affecting the results. The *R* package **robustHD** (Alfons, 2013) provides functionality for robust linear model selection with complex high-dimensional data. More specifically, the implemented functionality includes robust least angle regression (Khan et al., 2007) and sparse least trimmed squares regression (Alfons et al., 2013). The package follows a clear object-oriented design and takes advantage of *C++* code and parallel computing to reduce computing time. In addition, cross-validation functionality to select the final model, as well as diagnostic plots to evaluate the model selection procedures are available in **robustHD**.

References

Alfons, A. (2013). **robustHD***: Robust methods for high-dimensional data*. *R* package version 0.3.0.

Alfons, A., C. Croux, and S. Gelper (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*. In press.

Khan, J., S. Van Aelst, and R. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association 102*(480), 1289–1299.

HGLMMM and JHGLM: Package and codes for (joint)hierarchical generalized linear models

Marek Molas[1,⋆], **Maengseok Noh**[2], **Seungyoung Oh**[1], **Youngjo Lee**[1]

1. Department of Statistics, Seoul National University, Seoul, Korea
2. Department of Statistics, Pukyong National University, Busan, Korea

⋆Contact author: markmolas@yahoo.com

Keywords: hierarchical likelihood, random effects models

Recently, some *R* packages have been developed for fitting hierarchical generalized linear models (HGLMs) of Lee and Nelder (1996) by using h-likelihood procedures such as **HGLMMM** package. HGLMs were developed from a synthesis of generalized linear models, random-effect models and structured dispersion models. Joint hierarchical generalized linear models(JHGLMs) are a further extension of HGLMs , where several HGLM processes are linked together through multivariate Gaussian distribution. Package **HGLMMM** fits standard hierarchical generalized linear models, we also provide the routines for the analysis of JHGLMs.

Package **HGLMMM** can fit hierarchical generalized linear models, these are normally, gamma, Poisson or Binomial responses with conjugate Bayesian random effects distributions. It offers many various distributional assumptions for random elements of the models without referring to sampling mechanisms. Complex designed can be handled as cross-sectional, multilevel structures and cross over experiments. Further we offer R routines to fit several hierarchical generalized linear models at once, by assuming a multivariate Gaussian distribution for all the random effects in the system.

This presentation will introduce **HGLMMM** package and joint HGLM routines with adequate theory and practical applications.

Fitting regression models for polytomous data in R

Yiwen Zhang[1*], **Hua Zhou**[1]

1. Department of Statistics, North Carolina State University
*Contact author: yzhang31@ncsu.edu

Keywords: multivariate generalized linear models (MGLM), categorical data analysis, iteratively reweighted Poisson regression (IRPR), sparse regression, Nesterov method

Data with multivariate categorical responses frequently occur in modern applications, such as RNA-seq analysis, pattern recognition, document clustering, and image reconstruction. The commonly used multinomial-logit model is limiting due to its restrictive mean-variance structure. More flexible models such as Dirichlet-multinomial, generalized Dirichlet-multinomial, and negative multinomial regressions are needed to accommodate over-dispersion and more general correlation structures. Fitting these models are difficult though, partly due to the fact that they do not belong to the exponential family. We propose an iteratively reweighed Poisson regression (IRPR) algorithm for maximum likelihood estimation and an accelerated projected gradient method for variable selection by penalized regression. The methods are implemented in a comprehensive *R* package **MGLM**. Numerical results are demonstrated on both simulation study and disease mapping based on real RNA-seq data.

An exposé of naming conventions in R.

Rasmus Bååth[1⋆], **Martin Jönsson**[2]

1. Cognitive Science, Departement of Philosophy, Lund University, Sweden.
2. Departement of Mathematical Statistics, Lund University, Sweden.
⋆Contact author: rasmus.baath@lucs.lu.se

Keywords: *R* programming, Naming conventions, Coding Style

R is one of the most heterogeneous programming languages when it comes to naming conventions. An example of this is the conventions for function names where most other modern languages are divided between using underscore_separated and lowerCamelCase names and have official guidelines stating which convention to prefer. In the *R* community there are no less than five different naming conventions for function names in use and many unofficial guidelines disagreeing on which one to prefer. *R* is rapidly gaining in popularity and there is a steady stream of newcomers having to decide what naming conventions to adopt. If you are a newcomer to *R* or if you are a package developer you would probably want to adhere to the current naming conventions of the *R* community, but how to know what the most common convention is? While there are no official guidelines there fortunately exist ample information regarding what conventions are used *in practice* as the *Comprehensive R Archive Network* (http://cran.r-project.org/) contains the code and documentation of over 4000 R packages.

We downloaded the documentation of each package on CRAN and counted what proportion of function and parameter names that followed different naming conventions. A summary of the results were published in the *Programmer's Niche* section in the *R* journal Bååth (2012) with the main finding being that the most common naming convention for function names was lowerCamelCase (55% matched this convention) and that most argument names were period.separated (83 %). It was also found that many packages use mixed naming conventions (28 % of the packages mix three or more conventions). A lot of interesting findings did not make it into the *Programmer's Niche* article however. Using information regarding when a package was first released it is possible to study naming convention usage over time (see figure 1) which seems to be rather stable with a possible downward trend for using period.separated names. The CRAN documentation is also full of examples of function names that don't seem to follow any naming convention, for example, the function names `Balanced.Initialization` and `mv_truthTable` and parameter names `To.prime`, `missing.PlugIn` and `outputSGP_INDIVIDUAL.content_areas`. Except for how to handle identifiers consisting of many words there are also many other implicit naming conventions worth pointing out.

References

Bååth, R. (2012, December). The state of naming conventions in R. *The R Journal 4*(2), 65–73.

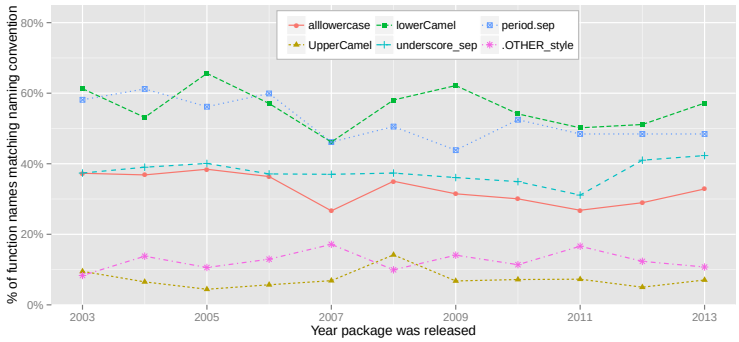


Figure 1: Naming convention usage on CRAN from 2003 to March 2013.

Statistical Machine Translation tools in *R*

Neda Daneshgar[1,2] , Majid Sarmad[1,3,*]

1. Ferdowsi University of Mashhad, Faculty of Mathematical Sciences
2. PhD Student in Statistics (Inference)
3. Assistant Professor in Statistics (Statistical Computations)

*Contact author: sarmad@um.ac.ir

Keywords: SMT, MPI, IBM Models

Statistical Machine Translation (SMT) is a method to work on parallel corpuses to translate a text in another language (as much as fluent) without any knowledge of both languages. The book is written by Koehn (University of Edinburgh) explains IBM models (including algorithms) to achieve the goal. There are some tools to work on parallel corpuses but no package in *R*. We have started to write the package in *R* and IBM model-I is roughly completed and works fine. The most important problem is that parallel corpuses are big (better to say huge), then we need to apply some techniques to handle the large objects with a reasonable speed. MPI (Message-Passing Interface) packages in *R* would help us to make the codes for a parallel processing. Furthermore, we are going to compare our speed with the other tools like *Moses* and *Giza++*.

Reference

Koehn. P. (2010). Statistical Machine Translation. Cambridge.

Reference classes: a case study with the poweRlaw package

Colin Gillespie

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
NE1 7RU
UK
Contact author: colin.gillespie@newcastle.ac.uk

Keywords: Reference classes, power-law distributions, efficiency, parallel computing

Power-law distributions have been used extensively to characterise many disparate scenarios, *inter alia*, the sizes of moon craters and annual incomes (Newman, 2005). Recently power-laws have even been used to characterize terrorist attacks and interstate wars (Cederman, 2003). However, for every correct characterisation that a particular process obeys a power-law, there are many systems that have been incorrectly labelled as being scale-free; see for example, Stumpf and Porter, 2012.

Part of the reason for incorrectly categorising systems with power-law properties is the lack of easy to use software. The **poweRlaw** package aims to tackles this problem by allowing multiple heavy tail distributions, to be fitted within a standard framework (Gillespie, 2013)). Within this package, different distributions are represented using reference classes. This enables a consistent interface to be constructed for plotting and parameter inference.

This talk will describe the advantages (and disadvantages) of using reference classes. In particular, how reference classes can be leveraged to allow fast, efficient computation via parameter caching. The talk will also touch upon potential difficulties such as combining reference classes with parallel computation.

References

Cederman, L.-E. (2003). Modeling the size of wars: from billiard balls to sandpiles. *American Political Science Review 97*(01), 135–150.

Gillespie, C. S. (2013). *poweRlaw: Fitting power laws to discrete and continuous data*. R package version 0.16.1.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics 46*(5), 323–351.

Stumpf, M. P. and M. A. Porter (2012). Critical truths about power laws. *Science 335*(6069), 665–666.

Combining R and Python for scientific computing

Felipe Ortega[1,⋆], **Javier M. Moguerza**[1] **and Emilio L. Cano**[1]

1. Universidad Rey Juan Carlos
⋆Contact author: felipe.ortega@urjc.es

Keywords: Scientific computing, Python, Django, lxml, programming interfaces

Over the past decade, *Python* has become one of the most popular programming languages for scientific computing. The large number of available libraries to extend its features, including **NumPy**, **SciPy**, as well as **Pandas** (for data mining) have led *Python* to play a central role in numerous scientific projects requiring support for mathematical and statistical operations. In the same way, the *R* programming language has evolved to become today a mature and solid resource for statistical computing, specially regarding the availability of hundreds of contributed packages to extend its core functionalities.

However, in many ocasions developers need to combine the advantages of these two powerful programming languages to address complex challenges in data science projects. On the one hand, *Python* offers more flexible and intuitive alternatives to implement code for data retrieval, parsing and preparation. On the other hand, once the data are ready for the analsys, *R* exhibits a richer collection of packages that already implement most of the statistical and data visualization features required in practical studies. In addition to this, sometimes it is required to provide a convenient and structured way to access data and results from these analyses through a structured and simple API, so that they can be imported by other software systems that will reuse them for different purposes, or will display them on third-party GUIs.

Hence, the purpose of this presentation is to provide a summary of current practical strategies to combine *R* and *Python* code in data analysis, following a structured, maintanable and pragmatic approach. Concrete examples taken from different application domains, including software engineering, the study of open online communities and optimization in energy systems, will illustrate these strategies. Furthermore, this presentation will put special emphasis in two specific Phython projects that can be of great utility for scientific computing developers:

- **lxml** lxml Development team (2013), one of the most advanced and feature-rich libraries for parsing *XML* code in *Python*.
- **Django** Django Software Foundation (2013), a high-level web development framework in *Python* that is specially suitable for fast and maintanable implementation of REST APIs Fielding (2000).

References

Django Software Foundation (2013). Django. https://www.djangoproject.com/.

Fielding, R. T. (2000). *REST: Architectural Styles and the Design of Network-based Software Architectures*. Doctoral dissertation, University of California, Irvine.

lxml Development team (2013). lxml - processing xml and html with python. http://lxml.de.

Shiny: Easy web applications in R

Joe Cheng

RStudio, Inc.
joe@rstudio.com

Keywords: web applications, interactive, reactive programming

The **shiny** package provides a framework that makes it easy for *R* users to create interactive web applications. It includes a library of easy-to-use input widgets like sliders, drop-downs, and text fields, and easily displays plots, tables, and summaries.

No knowledge of web technologies is necessary, but Shiny users who do know HTML and JavaScript can extend the framework with new types of input/output widgets and visual themes. These Shiny extensions can then be bundled into *R* packages for easy reuse by other Shiny users.

This talk will include an introduction to Shiny, a walkthrough of simple and complex Shiny applications, and speculation on future directions of the framework.

References

RStudio, Inc. (2012). Shiny home page, http://rstudio.com/shiny/.

rapport, an *R* report template system

Aleksandar Blagotić[1,2,⋆], **Gergely Daróczi**[3,4,5,⋆⋆]

1. MSc student at the Department of Psychology, University of Niš, Serbia
2. Web and R developer at Easystats Ltd, United Kingdom
3. Assistant lecturer at the Pázmány Péter Catholic University, Hungary
4. PhD student at the Corvinus University of Budapest, Hungary
5. Founder at Easystats Ltd, United Kingdom

Contact authors: ⋆alex@rapporter.net and ⋆⋆daroczig@rapporter.net

Keywords: template, report, reproducibility, web-development

rapport is an *R* package aimed at creating reproducible statistical report templates. The goal of this talk is to discuss **rapport**'s unique approach to report reproducibility through a blend of literate programming and template-based reporting, that allows the user to replicate his analysis against any suitable dataset, by means of a simple *R* command. We will explain the usage of *template-specific inputs* that one can match against the dataset variables or custom *R* objects in order to produce a report. The role of *Pandoc* document converter in **rapport** templates will be discussed alongside its importance in export of the rendered reports to plethora of external formats. We will unveil how **rapport** uses **brew**-style tags (`<% %>`) in order to evaluate *R* expressions and how is the output of the evaluated expressions "guessed" and displayed in an appropriate manner. **pander** backend will also be discussed, as it provides a robust cache engine, applies a uniform look to all the **graphics**, **lattice** or **ggplot2** plots and permits the manipulation of template parts via *R* control structures. **rapport**'s (dis)similarities with packages like **brew**, **Sweave** or **knitr** will be exposed. We will see what justifies the claim that "**rapport** was built with the Web in mind" and how does **rapport** fit into web-developer's daily routine in general.

References

Blagotić, Aleksandar and Daróczi, Gergely. (2013). **rapport**: a report templating system, URL http://cran.r-project.org/package=rapport

Daróczi, Gergely (2013). **pander**: An R Pandoc Writer, URL http://cran.r-project.org/package=pander

Horner, Jeffrey. (2011). brew: Templating Framework for Report Generation. R package version 1.0-6, URL http://CRAN.R-project.org/package=brew

Seamless C++ Integration with Rcpp Attributes

J.J. Allaire[1,⋆]

1. RStudio
⋆Contact author: jj@rstudio.com

Keywords: C++, Rcpp

The **Rcpp** (Eddelbuettel and François, 2013) package allows users to effortlessly pass rich objects between *R* and *C++* code. The initial use cases for **Rcpp** were using *C++* within *R* packages as well as embedding *R* in *C++* applications. However, users are also highly interested in working with *C++* interactively, a possibility afforded by the **inline** package (Sklyar et al., 2012).

Attributes are annotations added to *C++* source files to provide additional information to the compiler (Maurer and Wong, 2008). *Rcpp Attributes* (Allaire et al., 2013) are a new feature that provides a high-level syntax for declaring *C++* functions as callable from *R* and automatically generating the code required to invoke them. The motivation for attributes is several-fold:

1. Reduce the learning curve associated with using *C++* and *R* together
2. Eliminate boilerplate conversion and marshaling code wherever possible
3. Seamless use of *C++* within interactive *R* sessions
4. Unified syntax for interactive work and package development

Rcpp supports attributes to indicate that *C++* functions should be made available as *R* functions, as well as to optionally specify additional build dependencies. The *C++* file can then be interactively sourced into *R* using the `sourceCpp` function, which makes all of the exported *C++* functions immediately available to the interactive *R* session. As a result of eliminating both configuration and syntactic friction, the workflow for *C++* development in an interactive session now approximates that of *R* code: simply write a function and call it.

Attributes can also be used for package development via the `compileAttributes` function. This enables ad-hoc interactive work done with `sourceCpp` to be easily migrated into a package for broader distribution.

This talk will cover the motivation for and implementation of attributes as well as review many practical examples of their use.

References

Allaire, J., D. Eddelbuettel, and R. François (2013). *Rcpp Attributes*. R package version 0.10.3.

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

Eddelbuettel, D. and R. François (2013). *Rcpp: Seamless R and C++ Integration*. R package version 0.10.3.

Maurer, J. and M. Wong (2008). Towards support for attributes in C++ (revision 6). In *JTC1/SC22/WG21 - The C++ Standards Committee*. N2761=08-0271.

Sklyar, O., D. Murdoch, M. Smith, D. Eddelbuettel, and R. François (2012). *inline: Inline C, C++, Fortran function calls from R*. R package version 0.3.10.

The R Service Bus: New and Noteworthy

Tobias Verbeke[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: tobias.verbeke@openanalytics.eu

Keywords: R Service Bus, Automation, Enterprise Service Bus, Architect

The R Service Bus (RSB) (OpenAnalytics, 2013b) is a popular open source solution to automate statistical analyses using *R*. It is a swiss army knife that allows all kinds of software systems or end user applications to make use of R-based statistical functionality.

In this presentation we will demonstrate the R Service Bus at work, showing how easily one can turn an arbitrary R package into an R Service Bus application and how one can address the functionality via a web client, using folder monitoring, via e-mail or via API calls.

In the second part of the presentation we will list new features that were added for a use case where RSB had to handle 8500 requests per second. It concerns primarily new authentication backends, client side pooling for extremely fast processing and integration with the Architect IDE (OpenAnalytics, 2013a).

References

OpenAnalytics (2010–2013b). The R Service Bus. A communication middleware and work manager for R. http://www.openanalytics.eu/r-service-bus.

OpenAnalytics (2011–2013a). Architect. http://www.openanalytics.eu/architect.

Outliers in multivariate incomplete survey data

Beat Hulliger[1,⋆]

1. University of Northwestern Switzerland
⋆Contact author: beat.hulliger@fhnw.com

Keywords: Missing value, robust covariance, Mahalanobis distance, data depth

The distribution of quantitative survey data usually is far from multivariate normal. Skew and semi-continuous distributions, particularly zero-inflated distributions, occur often. In addition, survey data often contains missing values. The sample design and the non-response process must be taken into account also. All together this mix of problems make multivariate outlier detection difficult. Examples of surveys where these problems occur are most business surveys and some household surveys like the Survey for the Statistics of Income and Living Condition (SILC) of the European Union. Usually the detection of a an outlier must be followed by a treatment. Often imputations are used to enable the use of methods for the analysis of the data which rely on a complete and cleaned data set.

In sample surveys the definition of an outlier usually cannot rely on a parametric model. The outlier generating mechanisms may be depend on other variables. Therefore, (Béguin and Hulliger, 2008) introduced the notion of an outlier at random and (Hulliger and Schoch, 2013) discuss a full model of outlier generating mechanisms and the connections between missingness and outlyingness.

Two possible outlier detection-and-imputation procedures are investigated. The BACON-EEM algorithm introduced by (Béguin and Hulliger, 2008) detects outliers under the assumption of a multivariate normal distribution, usually after a suitable transformation. Subsequently an imputation based on the multivariate normal distribution and back-transformation is carried out. The epidemic algorithm (Béguin and Hulliger, 2004) is based on a type of data depth. It is run forward to detect outliers and backward to impute for outliers. It is more difficult to parametrise but does not assume multivariate normality.

Using the SILC universe and the simulation tools prepared in the project AMELI (Münnich et al., 2011) the algorithms and variants of them are evaluated with data that is close to reality. The evaluation uses the impact on important indicators of SILC. In particular the choice of the tuning constants of the algorithms is discussed. The algorithms are implemented in an experimental package of *R* called **modi**.

References

Béguin, C. and B. Hulliger (2004). Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society, Series A: Statistics in Society 167*(2), 275–294.

Béguin, C. and B. Hulliger (2008). The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Survey Methodology Vol. 34, No. 1*, 91–103.

Hulliger, B. and T. Schoch (2013). Multivariate outliers in incomplete survey data. *Survey Methodology*. submitted 2013.

Münnich, R., S. Zins, A. Alfons, C. Bruch, P. Filzmoser, M. Graf, B. Hulliger, J.-P. Kolb, R. Lehtonen, D. Lussmann, A. Meraner, M. Myrskyl, D. Nedyalkova, T. Schoch, M. Templ, M. Valaste, and A. Veijanen (2011). Policy recommendations and methodological report. Research Project Report WP10 – D10.1/D10.2, FP7-SSH-2007-217322 AMELI.

Use of R and LaTeX for periodical statistical publications

Matjaž Jeran[1,⋆]

1. Banka Slovenije / Bank of Slovenia, Eurosystem, Department of financial statistics
⋆Contact author: matjaz.jeran@bsi.si

Keywords: reproducible research, literate programming, statistical reports

This presentation will address the question whether *R*, LaTeX and literate programming are sufficiently capable to produce high quality professional publications like monthly bulletins of a central bank. We will discuss the required features of the software to make the job of publishing easier. To find this out, we will pose the following questions.

Can *R* read a lot of different data formats? Can *R* manipulate data? Can *R* and LaTeX print free text mixed with the dynamic results of statistical computation? How can we use *R* and LaTeX to format a table and to draw a graph? Can *R* be used for making thematic maps? Can we integrate a publication from many separate pieces, made in different departments into a single volume? Are there any flaws in these procedures and how to avoid them?

All these questions will be answered with yes and demonstrated by code snippets of R and LaTeX using the *R* package **knitr**. The demo example will show *R* scripts reading data from relational databases, flat files, *MS Excel* and *PC-Axis* files used by statistical offices. The example will show free texts with embedded results of statistical computations, table, graph and a thematic map as typical elements of a professional publication. Some of the packages for reading data are: **RODBC**, **XLConnect**, **xlsx**, **pxR** and **XML**. Essential packages for producing tables and graphs are **xtable**, **rwoldmap** and **maptools**.

The presentation is based on the experience producing some real central banking publications that are produced regularly. We will also discuss some details in publications which cannot be fully automatic, and where human gray cells and hands must be activated to do some polishing of the publication.

References

Banka Slovenije / Bank of Slovenia Eurosystem: http://www.bsi.si/

European central bank statistical data warehouse: http://sdw.ecb.europa.eu/

Eurostat Search Database:
http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

Portals of mapping authorities in Slovenia: http://www.gu.gov.si/ and
http://e-prostor.gov.si/

Earth maps: Natural Earth: http://www.naturalearthdata.com/

Global administrative areas: http://www.gadm.org/

INSPIRE: http://inspire-geoportal.ec.europa.eu/

KaspeR: http://www.gis.si/kasper/si/index.html

Solving Dynamic Macroeconomic Models with R

Giuseppe Bruno[1⋆]

1. Bank of Italy, Research Area.
⋆Contact author: giuseppe.bruno@bancaditalia.it

Keywords: Bellman equation, Dynamic Programming, fixed point.

Building and solving a macroeconomic model is one of the most important tasks facing economists working in the Research divisions of a Central Bank. Solving a dynamic macroeconomic model consists in the optimization of a given objective function subject to a series of constraints. The standard deterministic problem considers a representative household which maximizes his life-time utility subject to the economy's resource constraints:

$$\max_{\{c_t\}} \left(\sum_{t=1}^{\infty} \beta^t u(c_t) \right), \quad \text{with } \beta \in (0,1) \tag{1}$$

1. output is given by $y_t = f(k_t) = c_t + i_t$, where: $f(\cdot)$ is the production function, k_t is the capital stock, c_t is the consumption and i_t is the investment;
2. capital stock dynamics is given by: $k_{t+1} = i_t + (1-\delta)k_t$, where δ is the depreciation rate;

Problem (1) can be expressed recursively in term of the Bellman equation (see Heer and Maußner (2008)): $V(k) = \max_{\{k_{t+1}\}} [u(f(k) + (1-\delta)k - k_{t+1}) + \beta \cdot V(k_{t+1})]$, where $V(k)$ is the value function. This equation can furtherly be transformed in following way:

$$V^{n+1}(k) = \max_{\{k_{t+1}\}} [u(f(k) + (1-\delta)k - k_{t+1}) + \beta \cdot V^n(k_{t+1})] \tag{2}$$

Assuming the concavity of the utility $u(\cdot)$ and the production function $f(\cdot)$, iterating equation (2) will converge monotonically to its fixed point, the stationary value function $V^*(\cdot)$. This algorithm, dubbed value function iteration (**VFI**), is the easiest to code but also the slowest to run. Once obtained the value function we can derive the policy function $K' = h(K)$ which provides the optimal capital accumulation policy.

Although *R*, so far, is still less common in economics than languages such as *Matlab* or *Gauss*, and is slower than lower level languages such as C and Fortran, it is very easy to extend *R* vocabulary with users' written functions. For this research we developed a set of functions for the solution of the recursive Bellman equation in deterministic and stochastic scenarios. **VFI** algorithm features a convergence rate with linear order, the employment of the Policy function iteration **PFI** (see for example Richter et al. (2011) provides an algorithm featuring quadratic convergence rate. In the paper we show some numerical applications in solving an optimal growth model employing a set of *R* functions for both the **VFI**, the **PFI** methods and some of their performance-improvement modifications. Solution time with *R* is often slower than *Matlab*. Nonetheless we show how to get some improvement employing the bytecode produced with the **compiler** package.

References

Heer, B. and A. Maußner (2008). Value Function Iteration as a Solution Method for the Ramsey Model. *CESifo Working Paper* (2278).

Richter, A. W., N. A. Throckmorton, and T. Walker (2011). "Accuracy, Speed and Robustness of Policy Function Iteration". `http://auburn.edu/~awr0007/AWR_files/RTW_Numerical.pdf`.

packdep: network abstractions of CRAN and Bioconductor

Radhakrishnan Nagarajan[1*], Marco Scutari[2]

1. Division of Biomedical Informatics, Department of Biostatistics, University of Kentucky, Lexington, USA
2. UCL Genetics Institute, University College London, London, UK
*Contact author: rnagarajan@uky.edu

Keywords: Network Analysis, CRAN, Bioconductor

Objective: *The objective of* `packdep` *is to model the associations between the R packages in CRAN and Bioconductor as networks and subsequently investigate their topological/statistical properties.*
Rationale: *While it is clear that the number of R packages in CRAN and Bioconductor is growing steadily with time, and that packages evolve significantly during their development, little attention has been given to understanding the associations between them. Such an understanding can provide novel system-level insights that capture their intricate wiring patterns. They may also be useful in identifying critical packages whose perturbation can challenge the stability of CRAN and Bioconductor. While we present some preliminary findings, we expect* `packdep` *to mature into a useful surveillance tool for CRAN and Bioconductor.*

Network Abstraction of CRAN and Bioconductor: The exponential growth of R packages has largely been attributed to the active user community and special interest groups that span a spectrum of disciplines including Bioinformatics. Reuse of existing packages and/or functionalities minimizes redundancies and is characteristic of open-source communities where user contributions are voluntary. However, reuse introduces *dependencies/associations between packages* that warrant a detailed investigation, since user contributed packages are often subject to significant changes during their devleopment. In R, associations fall under three broad categories of decreasing significance: (*i*) `depends` (*ii*) `imports` and (*iii*) `suggests` and can be modeled as a *directed graph*.
CRAN: As expected, analysis of CRAN network generated using the category (`depends`) revealed the *core packages* to be highly connected and dominant mediators. Removing the core packages and repeating the exercise on the weakly connected network component revealed the degree and betweenness distribution to be positively skewed reflecting a few packages are highly connected and dominant mediating comprising the tail of the distribution. Such distributions implicitly reflect the inherent non-random association patterns in CRAN. Top ten packages with maximal impact on others (i.e. out-degree) were `mvtnorm, Rcpp, coda, sp, ggplot2, rgl, XML, plyr, igraph, rJava`. Statistical packages (`mvtnorm, sp, coda`), graphical packages (`ggplot2, rgl`) are routinely used across a number of disciplines while (`XML, Rcpp, rJava`) integrate R to other popular software environments. The high rank `igraph` being may be attributed to the increasing emphasis on system and network science. Top ten dominant mediators in CRAN consisted of `gplots, Deducer, RCurl, hdrcde, ROCR, spdep, fda, geiger, distr, ks`. These packages while not highly wired by themselves, they do facilitate cross-talk across the highly wired packages establishing their critical role.
Bioconductor: A similar analysis on Bioconductor revealed the highly connected packages (out-degree) to be `AnnotationDbi, Biobase, oligoClasses, oligo, org.Hs.eg.db, BiocGenerics, affy, BSgenome, IRanges, graph` and the dominant mediators to be `AnnotationDbi, oligo, Biobase, GenomicFeatures, oligoClasses, affy, ShortRead, Biostrings, Rsamtools`. The high ranks of the annotation packages (`AnnotationDbi, Biobase, GenomicFeatures`) may be attributed to their critical role in the integration across disitinct high-throughput molecular assays and the increasing emphasis on translational research (`org.Hs.eg.db`). The widesrepad use of microarrays and recent shift towards high-throughput sequencing and GWAS may also explain the high ranks of (`oligo, Biobase, oligoClasses, affy`) and (`BSGenome, ShortRead, RSamtools`). Unlike CRAN there was a significant overlap between highly wired packages and dominant mediators in Bioconductor.

A more detailed investigation of these network abstractions is currently under investigation.

The Beatles Genome Project: Cluster Analysis of Popular Music in R

Douglas Mason[1]

1. Harvard University and Twitter, Inc.
*Contact author: douglasmason@gmail.com

Keywords: Information Retrieval, Pandas, Heirarchical Clustering, Visualization, Exploratory Analysis

We present a front- and back-end system for storing and retrieving abstracted musical information which is capable of generating meaningful musical visualizations by linking into *R*'s rich graphical packages. In particular, we apply *R*'s hierarchical clustering package to show statistical phenomena occurring in a corpus of popular songs written by the Beatles. We find that chords and melodic rhythms fall into clear clusters which distinguish each song, indicating possible principles of composition.

In the case of this presentation, we have focused on the Rolling Stone list of the 100 Greatest Beatles Song, which have all been entered into the database and cross-checked against other transcriptions. Our software front-end, built on the popular *MediaWiki* package, divides the task of musical input into individual phrases that can be typed using an easy-to-understand textual musical language called TinyNotation, developed at MIT and enhanced for our purposes. Once a song has been stored in this format, it it can be retrieved and processed using the *Python* package Music21[1]. The returned analysis is then fed through the *Pandas* package[2] to interface with *R* and to plot results for the user.

Early analysis shows that two features of the Beatles corpus are especially prominent: the inclusion of non-diatonic chords (see de Clercq and Temperley[3]) and the tendency for musical phrases from within a song to start on the same beat of the measure. To examine the presence of non-diatonic chords, we describe each song as a feature vector space in which each dimension correlates to a different chord root and modality. Each root is normalized to the tonal center and the value of a song in each dimension corresponds to the percentage of the song which plays that chord. We then normalize our feature space by applying tf*idf measures to indicate the level of surprise.

The results from our hierarchical cluster analysis show that most songs in the corpus (~70%) strongly emphasize non-diatonic chords, and among those, the majority (~80%) exhibit the strong presence of only one. Moreover, song clusters span many albums and almost never include two songs from the same album. Both of these results suggest that the Beatles were deliberate in limiting the harmonic palette of each song and careful not to repeat that palette in the same album.

We also examined the onset of musical phrases against the bar using a similar approach. Our analysis indicates high-level clustering of songs into on-beat rhythms or off-beat rhythms. We also find strong clustering among individual beat onsets. Each cluster includes songs from many albums, although due to the limited dimensionality of beat values (8 possible beats and off-beats compared to 24 possible chords) it is common for multiple songs in the same album to cluster together. Like the chordal analysis, however, almost all songs emphasize only one beat value, indicating that the beat-onset of musical phrases acts like a musical fingerprint.

References

[1] Michael Scott Cuthbert and Christopher Ariza (2010). music21: A toolkit for computer-aided musicology and symbolic music data. *Proceedings of the International Symposium on Music Information Retrieval*, pages 637–42.
[2] W. McKinney (2012). *Python for Data Analysis*, O'Reilly Media
[3] Trevor de Clercq and David Temperley (2011). A corpus analysis of rock harmony. *Popular Music*, 30:47–70.

The secrets of inverse brogramming

Richie Cotton[1*]

1. TDX Group
*Contact author: richierocks@gmail.com

Keywords: style, introductory, sig, assertive

Brogramming is the art of looking good while coding. An even more important concept is that of inverse brogramming: the art of writing good looking code. This talk describes techniques for writing stylish, idiomatic *R* code; and ways to structure your code to make it easier to debug and reuse. The **sig** package for examining function signatures, and the **assertive** package for checking function inputs are discussed.

References

Cotton, Richard (2013). **sig** package, http://cran.r-project.org/web/packages/sig.
Cotton, Richard (2013). **assertive** package, http://cran.r-project.org/web/packages/assertive.

Mapping Hurricane Sandy Damage in New York City

Charles DiMaggio[1,2,*]

1. Columbia University Departments of Anesthesiology and Epidemiology
2. Center for Injury Epidemiology and Prevention at Columbia
*Contact author: cjd11@columbia.edu

Keywords: disaster response, mapping, epidemiology

Hurricane Sandy Maps for New York City

In late October, 2012, Hurricane (later Superstorm) Sandy caused extensive damage in the North East United States. In New York it was very much a tale of two cities, with some areas quickly back to their stride, and other areas continuing to limp along. Here is some epidemiologic surveillance in the form of mapping housing damage and power outages in New York City.

The following code maps US Federal Emergency Management Administration (FEMA) surveillance data on housing damage and inundation and power outages in New York City a week after the storm touched down. The FEMA data are available via Google Crisis Maps here. The New York City borough boundary files are available from Bytes of the Big Apple The outage data files are here and here . A full write up is here .

Preparing the Housing Data

```
library(maptools)
library(rgdal)

fema.points<-readOGR(".../femaPoints/", "femaPoints")
boros<-readOGR(".../nybb/", "nybb")

fema.points<-spTransform(fema.points, CRS("+proj=longlat +datum=NAD83"))
boros<-spTransform(boros, CRS("+proj=longlat +datum=NAD83"))
```

This syntax can be used for some simple plots.

```
plot(fema.points,col="red", pch=20, cex=.1)
plot(boros, add=T, lty=1, lwd=.5)
```

This syntax overlays (actually indexes) the FEMA data to NYC boundaries (see the vignette for an explanation).

```
 #vignette("over")
plot(fema.points[boros,], col="red", pch=20, cex=.3)
plot(boros, add=T, lty=1, lwd=.2)
title(main="Hurricane Sandy Housing Damage",
sub="FEMA Flyover Data November 2012")
```

Mapping Housing Damage

These maps are for all levels of damage. You can restrict to and explore the different levels of inundation and damage using indexing and the faceting capabilities of ggplot2.

Unlocking a national adult cardiac surgery audit registry with *R*

Graeme L. Hickey[1*], Stuart W. Grant[2], Ben Bridgewater[1,2]

1. Northwest Institute of BioHealth Informatics, University of Manchester, UK
2. Department of Cardiothoracic Surgery, University Hospital of South Manchester, UK
*Contact author: graeme.hickey@manchester.ac.uk

Keywords: Cardiac surgery, Healthcare registry, Audit, Performance

Following the Bristol Royal Infirmary heart scandal, the Society of Cardiothoracic Surgery in Great Britain & Ireland (SCTS) established a world-leading clinical registry to collect data on all adult cardiac surgery procedures. To date this registry contains >480,000 records and 163 fields. The data includes patient demographics, comorbidities and clinical measurements, cardiac and operative details, and post-operative outcomes. We will describe examples of how *R* has been used recently to interrogate the SCTS registry and run a national governance programme for performance monitoring.

Understanding the data is vital to making decisions. The SCTS have recently used the **googleVis** package by Gesmann and de Castillo (2011) to visualize hospital- and surgeon-level data longitudinally over time as Google Motion Charts (SCTS, 2013a). This can be used to interrogate, for example, the risk-adjusted mortality rate of healthcare providers, whilst gaining an understanding of the variation due to sample size or inherent natural variability. It can also be used to understand the multivariate relationships between data; for example is postoperative length-of-stay related to patient age and the number of operations performed by each hospital? This tool has already been the instigator of a number of clinical and care-quality investigations.

Monitoring performance of surgeons requires a broad portfolio of tools. First, statistical modelling tools, for example `glm` or `glmer`, are required to appropriately 'risk-adjust' outcomes. Second, functions to aggregate and summarize the data in different ways over healthcare providers are required. Finally, graphical tools are required to present the results as funnel plots and case mix charts to patients for scrutiny of their healthcare provision (SCTS 2013b).

"Real-world" databases are messy – the SCTS registry is no exception. Cleaning data can be complicated, especially if there interdependencies between data frame rows and columns. Synonyms and homonyms required homogenizing; numerical, temporal and clinical conflicts required resolving; and duplicate records required accurate identification and removal. Previously this was a terminal obstacle facing cardiac surgeons in their bid to unlock the potential of this data. A registry-specific *R* package has been written to fully automate the cleaning in a transparent and reproducible manner, thus enabling analyses of the data.

References

Gesmann M and de Castillo D (2011). googleVis: Interface between R and the Google Visualisation API. *The R Journal 3*, 40-44.

SCTS (2013a). Dynamic charts, http://www.scts.org/DynamicCharts.

SCTS (2013b). Performance reports, http://www.scts.org/patients/default.aspx.

Renjin: A new *R* interpreter built on the JVM

Alexander Bertram[1*]

1. BeDataDriven
*Contact author: alex@bedatadriven.com

Keywords: JVM, Optimization, Big Data

Renjin is a new interpreter for the *R* language built on the Java Virtual Machine. It is intended to be 100% compatible with GNU R, to run on Cloud-based PaaS' like Google AppEngine, and generally to open new opportunities for embedding libraries and programs written in *R* within larger systems. Though still under development, the interpreter is already used in production by several companies to embed existing R packages within web applications.

In addition to facilitating integration, Renjin's principal advantage over GNU R is an additional layer of abstraction between *R*-language code and *R* data structures. Renjin allows developers to provide their own implementations of *R* vectors, so that the same *R*-language code might compute on an in-memory array, a memory-mapped file, or a rolling buffer over a database cursor.

This layer of abstraction allows other optimizations within the interpreter itself, such as “views” on large vectors. In Renjin, most primitive operations will not allocate new memory for a computation on a large vector, but simply return a new view of the underlying data. This makes it possible to defer computation until late in a program at which point it is compiled to optimized JVM bytecode in such a way that takes the entire computation, and memory limits, into account.

There are naturally many other initiatives to improve *R*'s handling of large data, but Renjin has the distinct advantage of requiring no change to existing *R*-language code: the underlying storage of the data is invisible to the statistician.

The presentation will walk though a few concrete examples of how Renjin has helped solve problems in production, and lay out future paths for development.

References

Bertram, Alexander (2013). JVM-based Interpreter for the R Language for Statistical Computing, http://code.google.com/p/renjin

Using Lazy–Evaluation to build the G.U.I.

Jorge Luis Ojeda Cabrera[1,⋆]

1. Dep. Mtodos Estadsticos, U. de Zaragoza
⋆Contact author: jojeda@unizar.es

Keywords: G.U.I. development, Lazy–Evaluation, Web interface, Computation on the Language, Functional Programming.

The aim of this work is to introduce and discuss a G.U.I. building approach that relies on the Lazy–Evaluation and Functional Programming capabilities the *R* enjoys. The main idea behind this approach is to separate G.U.I. building code from the proper statistical computation code through the Computation on the Language features exhibited by *R*. In this way, this strategy makes it easy for both, statistical code developer and the statistical user to interact with code. Furthermore, because of the flexibility this G.U.I. building strategy enjoys, the user interaction experience can be easily deployed in several different user interfaces frameworks as the Web or a Desktop with minimal coding costs in an elegant and efficient way.

Package **miniGUI** implements this strategy allowing simple, ready to use and fully customized interaction with functions coded in *R* by means of the **tcltk** package. It uses *R* capabilities to compute on the Language jointly with Lazy–Evaluation to access to the code of a function `f` mapping its arguments to a **tcltk** widget that, once displayed, grants the execution of the function code.

As it is also shown in this work by means of the C.G.I. machinery provided by **FastRweb**, the same strategy can be used to enable Web interaction with code at minimal coding costs, except possibly for that code related to the graphical customization of the Web pages. This is due to the fact this strategy is based on simple *R* functions to code the the way the user inputs the arguments, so simple function coding leads different user input methods.

References

Cabrera, J. L. O. (2012). *miniGUI: tktcl quick and simple function GUI.* R package version 0.8.0.

Dalgaard, P. (2001, September). A primer on the R-Tcl/Tk package. *R News 1*(3), 27–31.

Hudak, P. (1989, September). Conception, evolution, and application of functional programming languages. *ACM Comput. Surv. 21*(3), 359–411.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Urbanek, S. (2008). Fastrweb: Fast interactive web framework for data mining using r. http://urbanek.info/research/pub/urbanek-iasc08.pdf.

Urbanek, S. and J. Horner (2012). *FastRWeb: Fast Interactive Framework for Web Scripting Using R*. R package version 1.1-0.

Survo for R - Interface for Creative Processing of Text and Numerical Data

Reijo Sund[1,⋆]

1. National Institute for Health and Welfare (THL), Helsinki, Finland
⋆Contact author: reijo.sund@helsinki.fi

Keywords: statistical software, text editor, user interface

Nowadays most statistical software packages contain interfaces to *R* in order to offer extended functionality for their users. We demonstrate a contrasting approach in which the functionality of an integrated environment for statistical computing and related areas known as *Survo* has been fully incorporated into *R* as an open source package **muste** freely available from the R-Forge repository.

Survo has been developed since the early 1960s by professor Seppo Mustonen. First full implementation for a statistical programming system *Survo* was *SURVO 66* on Elliott 803 and 503 computers. Honeywell H1640-series implementation was known by the name of *SURVO/71*. The next generation of Survo was the *SURVO 76* system working on the Wang 2200 minicomputer. The PC-version *SURVO 84* and especially the following *SURVO* 84C version grow to remind more like a statistical operating system than just a program for statistical analyses. *SURVO 98* expanded the software to work in 32-bit systems and *SURVO MM* in the Windows environment. Multiplatform open source *R* package version has been developed under the name of *Muste*. It is a sophisticated mixture of Mustonen's original *C* sources and rewritten I/O functions that utilize *R* and *Tcl/Tk* extensively.

Features of *Survo* include file-based data operations, flexible data preprocessing and manipulation tools, various plotting and printing facilities, a teaching friendly matrix interpreter, so-called editorial arithmetics for instant calculations, a powerful macro language, plenty of statistical modules and an innovative text editor based GUI (originally invented in 1979) that allows to freely mix commands, data and natural text encouraging towards reproducible research with ideas similar to literate programming. In addition, several properties have been developed to make the interplay with *R* from the GUI seamless.

By giving direct access to these additional useful features of *Survo* for every *R* user, the options available for data processing and analysis as well as teaching within *R* are significantly expanded. Text editor based user interface suits well for interactive use and offers flexible tools to deal with issues that may be challenging to approach using standard *R* programming. This new version of *Survo* also concretely shows how it is technically feasible to implement a whole full-featured statistical software within another statistical software.

References

Alanko T, Mustonen S, Tienari M (1968). A Statistical Programming Language SURVO 66. *BIT 8*, 69–85. http://dx.doi.org/10.1007/BF01939330

Mustonen S (1992) Editorial interface in Statistical Computing and Related Areas. In *COMPSTAT 1992, 10^{th} Conf. on Computational Statistics, (Neuchâtel, Switzerland)*, pp. 17–32. http://www.survo.fi/publications/COMPSTAT_1992.pdf

Sund R, Vehkalahti K, Mustonen S (2012). Muste - editorial computing environment within R. In *COMPSTAT 2012, 20^{th} Conf. on Computational Statistics, (Limassol, Cyprus)*, pp. 777–788. http://www.survo.fi/muste/publications/sundetal2012.pdf

Survo Systems (2013). Survo homepage, http://www.survo.fi/english/.

Using R in teaching statistics, quality improvement and intelligent decision support at Kielce University of Technology

Zdzisław Piasta[1,*]

1. Faculty of Management and Computer Modelling, Kielce University of Technology, Poland
*Contact author: piasta@tu.kielce.pl

Keywords: teaching, statistics education, machine learning, *R Data Miner*, *R Commander*

For many years the author has been giving the statistics courses for students studying engineering and management at Kielce University of Technology. More advanced courses for MSc and PhD students concern methods of quality improvement with the SPC and DoE approach. The courses on intelligent decision support are based on data mining and machine learning techniques. In all these courses data analysis software is a very important element of the teaching process.

In 1993 the author started to apply the SAS System in teaching. However, because of licence limitations and hardware requirements, only some of students were able to install the software on their PC. In turn, the open source machine learning software WEKA is not useful for the elementary courses of statistics. For those courses an attractive proposal was the free WebStat software. Its current successor, StatCrunch, offers useful statistical and graphical tools for elementary statistics courses, however students have to pay for the licence.

Last year the author decided to promote using R in teaching statistics, quality improvement and intelligent decision support. All students are able to install easily R software on their computers. The package **Rcmdr** give them a user-friendly GUI *R Commander* for data management, statistical data analysis and data visualization. They can also find many useful plug-ins to the *R Commander*. The package **rattle** offers GUI for more advanced data modeling, data mining and machine learning. All these tools are very helpful in understanding methods and techniques presented during lectures. Students apply these tools working on projects and theses. Some of them decided to prepare presentations for useR!2013 Conference. References contain several textbooks and manuals, useful in the process of teaching and applying R.

At the conference presentation the author will share his insights and experience in the use of R in teaching statistics, quality improvement and intelligent decision support.

References

G. Williams (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Springer.

T. Hastie, R. Tibshirani, J. Friedman (2009). Elements of Statistical Learning, http://www-stat.stanford.edu/ElemStatLearn.

P. Biecek (2011). Guide to the Package R, GiS, Wrocław (in Polish).

M. Walesiak, R. Gatnar (2009), Statistical Data Analysis with R, PWN, Warsaw (in Polish).

K. Kopczewska, T. Kopczewski, P. Wójcik (2009). Quantitative Methods in R, CeDeWu,Warsaw (in Polish).

J. Ćwik, J. Mielniczuk (2009). Statistical Learning Systems : R-based Exercises, OW PW, Warsaw (in Polish)

P. Biecek, Across Data Mining with R, http://www.biecek.pl/R/naPrzelajPrzezDM.pdf. (in Polish)

Facilitating genetic map construction at large scales in R

Emma Huang[1,*], Rohan Shah[1], Andrew George[1], Colin Cavanagh[2]

1. CSIRO Mathematics, Informatics and Statistics and Food Futures National Research Flagship, Dutton Park, QLD Australia
2. CSIRO Plant Industry and Food Futures National Research Flagship, Acton, ACT Australia
*Contact author: emma.huang@csiro.au

Keywords: GPU, MAGIC, RIL, linkage, recombination

High-throughput low-cost genotyping has made feasible the construction of high-density genetic maps in many species. To take advantage of this, new multi-parent experimental designs are becoming popular, as they have increased resolution and genetic diversity relative to traditional designs. However, the increased complexity of these designs necessitates novel computational approaches to map construction to deal with the high density of genetic markers.

We have developed the *R* package **mpMap** [1] to address the issues of genetic map construction at large scales in multi-parent experimental cross designs. To deal with increasing quantities of genetic data, we have integrated *R* code with both *C* and *CUDA* code to parallelize computational tasks. Functions for genetic map construction automatically take advantage of multiple cores within a machine or Graphics Processing Units, and can also be distributed to multiple cores. Functions for interactive visualization allow for fast manual checking of results from automated steps in the process.

We will step through the map construction process using this package, and the results from applying it to a four-parent recombinant inbred line cross.

References

1. Huang BE and George AW (2011). *R*/mpMap: A computational platform for the genetic analysis of multi-parent recombinant inbred lines. *Bioinformatics* 27, 727-729.

Elevating R to Supercomputers

Drew Schmidt[1,*], Wei-Chen Chen[2], Pragneshkumar Patel[1], George Ostrouchov[1,2]

1. Remote Data Analysis and Visualization Center — University of Tennessee Knoxville, USA
2. Computer Science and Mathematics Division — Oak Ridge National Laboratory, USA
*Contact author: schmidt@math.utk.edu

Keywords: parallel computing, high performance computing, big data, SPMD

The biggest supercomputing platforms in the world are distributed memory machines, but the overwhelming majority of the development for parallel R infrastructure has been devoted to small shared memory machines. Additionally, most of this development focuses on task parallelism, rather than data parallelism. But as big data analytics becomes ever more attractive to both users and developers, it becomes increasingly necessary for R to add distributed computing infrastructure to support this kind of big data analytics which utilize large distributed resources.

The *Programming with Big Data in R* (pbdR) project aims to provide such infrastructure, elevating the R language to these massive-scale computing platforms. The main goal of the project is to empower data scientists by bringing flexibility and a big analytics toolbox to big data challenges, with an emphasis on productivity, portability, and performance. We achieve this in part by mapping high-level programming syntax to portable, high-performance, scalable, parallel libraries such as MPI and ScaLAPACK. This not only benefits the R community by enabling analysis of larger data than ever before with R, but it also benefits the supercomputing community which, to date, mostly only dabbles superficially with statistical techniques.

A major focus of the project is ease of use, with great effort spent towards minimizing the burdens of supercomputing for R users and developers. Programs written using pbdR packages are written in the Single Program/Multiple Data, or SPMD style (not to be confused with SIMD architecture computers), which is a very natural extension of serial programming for distributed platforms. This paradigm together with extensive use of R's S4 methods allows us to create highly scalable tools with nearly-native serial R syntax.

In this talk, we will discuss some of the early successes of the pbdR project, benchmarks, challenges, and future plans.

References

Blackford, L. S., J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley (1997). *ScaLAPACK Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Chen, W.-C., G. Ostrouchov, D. Schmidt, P. Patel, and H. Yu (2012). pbdMPI: Programming with big data – interface to MPI. R Package, URL http://cran.r-project.org/package=pbdMPI.

Gropp, W., E. Lusk, and A. Skjellum (1994). *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. Cambridge, MA, USA: MIT Press Scientific And Engineering Computation Series.

Ostrouchov, G., W.-C. Chen, D. Schmidt, and P. Patel (2012). Programming with big data in R. URL http://r-pbd.org/.

Schmidt, D., W.-C. Chen, G. Ostrouchov, and P. Patel (2012). pbdDMAT: Programming with big data – distributed matrix algebra computation. R Package, URL http://cran.r-project.org/package=pbdDMAT.

R in Java: Why and How?

Tomas Kalibera*, Petr Maj, Jan Vitek

Purdue University, West Lafayette, IN, USA
*Contact author: kalibera@cs.purdue.edu

Keywords: R Language Runtime, Java, R Performance

As *R* is becoming increasingly more popular and widely used, two great challenges have emerged: performance and big data.

Increasingly more computation time is being spent in the *R* code as opposed to the numerical libraries. This has performance penalties and analysts often end up rewriting the hot spots of their *R* code to *C/C++*, which is time consuming and error prone. The current implementation of *R* has been around for about 2 decades (some parts of it nearly 4) and it would be very hard to extend it with today's state-of-the-art optimizations such as those present in the *C/C++* compilers. With hot spots being in the *R* code, the lack of parallelism in the *R* code is becoming a performance issue as well: current multi-core systems cannot be efficiently employed. Adding multi-threading to the *R* language would be hard within the current implementation.

R is being used for increasingly larger data. The data size limitations imposed by the use of 32-bit integers in the present *R* interpreter for encoding vector offsets are becoming a bottleneck on todays machines with large amounts of RAM. Data analysis these days and in the near future, however, needs to be done also on much larger data that would ever fit onto a single machine. Such data is typically stored in a cluster/cloud, often heavily cached in RAM of many nodes or even fully included in RAM of the nodes. Could *R* be made run in the cloud, evaluating parts of *R* expressions on the nodes where the data is?

We aim to attack these problems with a new *R* engine built on top of a *Java* virtual machine. The benefits we get from *Java* are good integrated support for multi-threading, a modern garbage collector, and a better integration with the cloud and databases. Choosing *Java* instead of say *C++* brings also a number of challenges. A big challenge is accessing well proven numerical libraries implemented in *C/Fortran*, such as LAPACK/BLAS, but also the Rmath library and other numerical codes present in *R*. Accessing them from *Java* incurs installation burden and for short-running operations has a performance overhead. Converting them to *Java* is difficult and the resulting code is likely to be slower for large data, as has been reported for the automatically converted codes of LAPACK/BLAS. A similar challenge is the use of *R* packages, parts of which are again implemented in *C* or *Fortran*.

We will explain the status of the project, FastR, currently on small benchmarks. On these we have seen speedups between 2x and 15x over the latest version of the *R* interpreter. We will provide some thoughts about where to go from there.

Rhpc: A package for High-Performance Computing

Ei-ji Nakama[1], **Junji Nakano**[2⋆]

1. COM-ONE Ltd. Ishikawa, Japan
2. The Institute of Statistical Mathematics, Tokyo, Japan
⋆Contact author: nakanoj@ism.ac.jp

Keywords: BLAS, CPU affinity, MPI, Supercomputer

Packages **snow** and **Rmpi** are usually used for realizing high performance computing in *R*. Package **snow** provides parallel processing functions by using several low level parallel computing mechanisms including MPI, which is mainly used for high performance computing in supercomputers. Although **snow** and **Rmpi** are useful and reliable packages, we need some additional functionalities for using recent supercomputer hardware sufficiently.

Package **Rhpc** is implemented for improving **snow** with **Rmpi** in some sense. It provides `clusterExport`, `clusterCall` and `clusterAapply` like functions implemented in *C* to use MPI functions directly and efficiently. It also provides CPU affinity setting functions to be used with OpenMP functions and optimized BLAS libraries such as GotoBLAS. These functions can reduce the latency of parallel computing in *R*.

References

L. Tierney, A. J. Rossini, N. Li, and H. Sevcikova (2013). CRAN - Package snow, http://cran.r-project.org/web/packages/snow/.

H. Yu (2013). CRAN - Package Rmpi, http://cran.r-project.org/web/packages/Rmpi/.

DCchoice: a package for analyzing dichotomous choice contingent valuation data

Tomoaki Nakatani[1⋆]

1. Department of Agricultural Economics, Hokkaido University
⋆Contact author: naktom2@gmail.com

Keywords: Dichotomous Choice, Contingent Valuation, Willingness to Pay

The **DCchoice** package provides functionalities for the analysis of data collected from dichotomous choice contingent valuation (CV) surveys. CV surveys study consumers' willingness to pay (WTP) for non-market goods. The areas of application range across environmental, agricultural, and health economics, to name a few. Two methods are used to ask for a respondent's WTP. One is a single-bounded format, whereby a particular bid is suggested once to a respondent, who determines whether the suggested bid is accepted or not. Thus, the single-bounded format provides data on whether the respondent's true WTP lies above or below the suggested bid. The second method is a double-bounded format, whereby a first bid is presented to the respondent and, given the answer to the first bid, a higher or lower second bid is offered. Therefore, the double-bounded format clarifies the interval containing the respondent's true WTP.

Modeling strategies can be either parametric or nonparametric. Parametric models are based on the utility difference approach (Hanemann, 1984). Two functions are offered in **DCchoice**, `sbchoice()` for single-bounded data and `dbchoice()` for double-bounded data. Both functions have an optional argument for specifying either a logistic, log-logistic, normal, or log-normal error distribution. Nonparametric models are based on the Kaplan–Meier–Turnbull approach, and can be applied using the command `turnbull.sb()` for single-bounded data or `turnbull.db()` for double-bounded data. For single-bounded data, the method of Kriström (1990) is also applicable through `kristrom()`. Empirical survival functions can be plotted by the generic function `plot()`.

Under the kind permission of the original authors, **DCchoice** includes example data sets from previous research, such as Kriström (1990), Carson et al. (1992), and Whitehead (1995). This enables the user to (at least partially) reproduce their outcomes.

In this presentation, the basic usage of **DCchoice** will be illustrated by analyzing example data.

References

Carson, R. T., R. C. Mitchell, W. M. Hanemann, R. J. Kopp, S. Presser, and P. A. Ruud (1992). *A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill*. Report to the Attorney General of the State of Alaska. Natural Resource Damage Assessment Inc.

Hanemann, W. M. (1984). Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics 66*(2), 332–341.

Kriström, B. (1990). A non-parametric approach to the estimation of welfare measures in discrete response valuation studies. *Land Economics 66*(2), 135–139.

Whitehead, J. C. (1995). Willingness to pay for quality improvements: Comparative statics and interpretation of contingent valuation results. *Land Economics 71*(2), 207–215.

Systems biology: modeling network dynamics in R

Aditya M. Bhagwat[1]

1. OpenAnalytics, BVBA
*Contact author: aditya.bhagwat@openanalytics.eu

Keywords: systems biology, network dynamics, dynamic modeling, kinetics

Systems biology is an exciting field, covering many different scientific activities. In this talk, I will zoom into one particular such activity: the study of network dynamics. I will discuss how a network of interacting components can be formulated, how appropriate kinetics can be defined, how the network can then be translated into a system of differential equations, and how this system can be integrated to study its dynamic behaviour using the *R* package **deSolve**. Finally, I will discuss how models can be exchanged through the SBML format using the *R* package **rsbml**.

References

Karline Soetaert, Thomas Petzoldt, R. Woodrow Setzer (2010). Solving Differential Equations in R: Package deSolve Journal of Statistical Software, 33(9), 1--25. URL http://www.jstatsoft.org/v33/i09/.

Evolutionary multi-objective optimization with R

Ching-Shih Tsou[1,*]

1. Institute of Information and Decision Sciences, National Taipei College of Business, Taipei 10051, Taiwan
*Contact author: cstsou@mail.ntcb.edu.tw

Keywords: multi-objective optimization, evolutionary algorithms, non-dominated solutions

Optimization is a process of finding and comparing feasible solutions until no better solution can be found. Evolutionary algorithms can find multiple optimal solutions in one single simulation run due to their population approach. This characteristic is especially important for multi-objective optimization because the goal of generating the Pareto front is itself bi-objective of convergence to the true front while acieving a well-distributed set of solutions (Deb, 2001). Several Evolutionary Multi-Objective Optimizers (EMOOs) have become popular in searching for the Pareto-optimal solutions. Two of them, Nondominated Sorting Genetic Algorithm (NSGA-II) and Strength Pareto Evolutionary Algorithm (SPEA2), are coded in R completely in this project (Deb *et al*., 2002, Zitzler *et al*., 2002). Common functions about the genetic operators and non-dominated solutions finding, such as tournament selection, bounded simulated binary crossover (SBX), bounded polynomial mutation, and continuously updated approach are presented first. Specific functions for above two EMOOs (non-dominated solutions sorting and crowding distance for NSGA-II / density estimation based on kth nearest neighbor and truncation operator for SPEA 2) are provided subsequently. Computational results against ZDT1 to ZDT 6 test problems (Deb, 2001) show that our implementations can correctly find the front in a reasonable time. Other EMOOs, speed up the R code and upload the EMOO package will be our next concerns. We believe that coding EMOOs algorithms in R is urgently needed and beneficial to the R community.

References

Deb, K. (2001). *Multi-objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, pp.33-43.

Deb, K., Pratap, A., Agrawal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197.

Zitzler, E., Laumanns, M., and Thiele, L. (2002). SPEA2 : improving the strength Pareto evolutionary algorithm for multiobjective optimization. *Evolutionary Methods for Design, Optimisation and Control, Giannakoglou, K., Tsahalis, D., Periaux, J., Papailiou, K., and Forgarty, T. (Eds.)*.

An integrated Solver Manager: using R and Python for energy systems optimization

Emilio L. Cano[1,*], **Antonio Alonso-Ayuso**[1], **Javier M. Moguerza**[1] **and Felipe Ortega**[1]

1. Universidad Rey Juan Carlos
*Contact author: emilio.lopez@urjc.es

Keywords: Optimization, interfaces, decision support systems, energy systems planning, risk management

EnRiMa (Energy Efficiency and Risk Management in Public Buildings) is an EU FP7 research project in which a Decision Support System (DSS), aiming at supporting building operators on both operational and strategic decisions, is being developed. Such DSS is composed of several integrated modules, which are in charge of specific tasks as a distributed system. The EnRiMa DSS relies on Stochastic Optimization as a framework for decision making under uncertainty. This approach provides optimal strategic decisions given all the scenarios considered, rather than for parameter estimates. Hence, not only average values of crucial parameters such as demand or investment and energy costs are used, but also their variability. That variability is implemented through the use of scenario trees. A scenario tree is a discretized representation of the stochastic process of the system.

The so-called Solver Manager module gathers the input from the rest of the modules through an interface, generates the problem instance, calls the optimization software, and delivers the solution eventually presented to the decision maker. Thus, the Graphical User Interface (GUI) DSS module provides input from and shows the solution to the user. The Scenario Generator tool DSS module provides the scenario tree structure and stochasticity information about the parameters. The DSS Kernel module provides data services to the rest of the modules, allowing a sound integration. The Solver Manager consists of two independent components: the interface and the "core script". The Solver Manager Interface allows us to separate communication tasks and other interaction features from the core features of the Solver Manager. The Solver Manager Interface has being built using the *Python* programming language. Using the data services provided by the DSS Kernel, the Solver Manager Interface retrieves the stochastic optimization problem instance data from a *MySql* database and creates XML files. Next, `data.frame` objects are created using the **XML** *R* package. The Solver Manager main script uses the **optimr** *R* library, which is being developed by the Department of Statistics and Operations Research at Universidad Rey Juan Carlos, and the **gdxrrw** package to interact with the GAMS optimization software. That script, after checking the model and the instance, generates the data for the optimization software, calls the optimizer and manages the output, storing the results in *R* objects ready for the Solver Manager Interface to make them available for the decision maker. Further development of the DSS will include the use of other optimizers, including *R* capabilities and APIs.

References

Energy efficiency and risk management in public buildings – EnRiMa. http://www.enrima-project.eu.

Conejo, A., M. Carrión, and J. Morales (2010). *Decision Making Under Uncertainty in Electricity Markets*. International Series in Operations Research and Management Science Series. Springer.

Dirkse, S. and R. Jain (2012). gdxrrw: An interface between GAMS and R. R package version 0.2.0.

Kaut, M., K. T. Midthun, A. S. Werner, A. Tomasgard, L. Hellemo, and M. Fodstad (2012). Dual-level scenario trees – scenario generation and applications in energy planning. Optimization Online. report 2012/08/3551.

Radar data acquisition, analysis and visualization using reproducible research with Sweave

Vladimir Skvortsov[1,2,⋆], **Keun Myeong Lee**[2]

1. Department of Computer Science, SUNY (The State University of New York) Korea
2. CEWIT (Center of Excellence in Wireless and Information Technology) Korea
⋆Contact author: vlad@sunykorea.ac.kr or vlad@cewit.re.kr

Keywords: radar, visualization, reproducible, Sweave, Fourier transform

Earlier we have proposed a novel radar-based inexpensive security or surveillance system for various applications as that presented in Skvortsov et al. (2012). The system comprises a radar network that creates a virtual fence or barrier that is invisible, and in case of intrusion it will give an alarm, determine the location of a target and target's velocity. The main focus of this study is on creating a software for working prototype of the radar unit. The system includes commercially available automotive radar, open source embedded hardware and software with signal processing routines. The radar unit is an automotive one-channel K-band frequency-modulated continuous-wave (FMCW) radar front-end and control board with serial interface.

The software application for the radar is built around serial terminal application and has two versions: *C++* with Qt library and *R*. The *R* version is used as a high level test lab for all algorithms which, in case of successful result, then are ported to *C++* for higher execution speed on embedded hardware. There are several choices for serial data acquisition in *R*. After comparison and due to some software/hardware requirements, we decided to proceed with method using low-level interface to *Java* VM with *R* package **rJava** and RXTX native library providing serial communication.

The FMCW radar linearly modulates a transmitted frequency signal. The received signal is then mixed with the emitted signal and finally results in an IF (intermediate frequency) signal. The Fourier Transform (FT) of the sampled signal is used to extract the distances to the different reflectors. The FT is crucial part of the radar data analysis, visualization and performance evaluation. To get a control over the resolution of range measurements, the analysis of several post-processing algorithms for target detection was carried out. We use a fast Fourier Transform in the `fft` function of *R* **stats** package and compare it with FFTW (Fastest Fourier Transform in the West), high resolution Fourier Transform (HRFT) and chirp Z transform (CZT). Our alarm principle implementation is based on perceptual (robust) hash algorithm for storage of current signal 'image' in memory and comparison of the signal and 'no-intrusion' signal. The `heatmap` function of *R* **stats** package is used for visualization of 'binary' signal output.

Scientific visualization has to be precisely controllable. The radar analysis plots are enhanced and customized with low level controls and optional arguments of both generic `plot` function of *R* **graphics** package and the `heatmap`. The standard `Sweave` figure including does not fit to our requirements. The optimal approach is to manually create the required graphics with `Sweave` options such as echo=false, results=hide and save it to a file, and then insert the file. In addition we have a control over how fonts are embedded.

Extensive experimental measurements confirm the advantages of the reproducible research approach, show the performance of the radar and reveal some directions for further developments. The routines created within the scope of this research may form a new radar analysis package in the future.

References

Skvortsov, V., K. M. Lee, and S. E. Yang (2012). Inexpensive Radar-Based Surveillance: Experimental Study. In *9th International Conference & Expo on Emerging Technologies for a Smarter World - CEWIT2012 (Songdo, Incheon, Korea)*.

Network Visualizations of Statistical Relationships and Structural Equation Models

Sacha Epskamp[1,⋆]

1. University of Amsterdam: Department of Psychological Methods
⋆Contact author: mail@sachaepskamp.com

Keywords: Data visualization, networks, psychometrics, structural equation modeling

I introduce the **qgraph** package (Epskamp et al., 2012) for *R*, which provides novel visualization techniques, using networks, for statistical relationships between variables in general and correctional structures in particular. For instance, a correlation matrix can be represented as a network in which each variable is a node and each correlation an edge; by varying the width of the edges according to the strength of the association, the structure of the correlation matrix can be visualized. This technique has many applications, such as allowing a researcher to detect complex structures in a dataset, validating the measurement model of a test and comparing individuals on differences in the correlation structure of repeated measures.

Extending on the back-end provided by **qgraph** the **semPlot** package (Epskamp, 2013) can be used to visualize path diagrams, parameter estimates and implied and observed correlations of structural equation models (SEM). This *R* package can import the output of several popular SEM packages including **Lisrel**, **Mplus** and *R* packages **sem**, **Lavaan** and **openMx.** Finally, **semPlot** also provides a bridge between these packages, allowing users to construct input for one SEM package based on the output of another.

References

Epskamp, S. (2013). *semPlot: Path diagrams and visual analysis of various SEM packages' output*. R package version 0.3.2.

Epskamp, S., A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software 48*(4), 1–18.

tableR
An R based approach for creating table reports from surveys

Oliver Bracht[1,*]

1. eoda
*Contact author: oliver.bracht@eoda.de

Keywords: Table Reports, Social Science, Market Research

Creating table reports from surveys is especially in the area of social science and market research a crucial task. There are software solutions on the market addressing this issue, but there is no satisfying R implantation yet.

tableR is both a stand-alone R-package as well as a Graphical User Interface for business users and non-programmers. It consists of three modules – questionnaire design, tabulation and graphics – which seamlessly integrate, but which can also be used independently.

tableR is based on an XML-Structure which defines a questionnaire and its tabulation resp. graphical representation in the very same document. Once a questionnaire is designed, a basic table report of univariate analysis is already defined. The tables can easily be extend by grouping and combining variables, adding totals and subtotals, titles and subtitles etc. The same principle holds true for graphics. Once the tables are designed, a basic graphic report is defined, which can be extended likewise.

As business users are very keen on (editable) Microsoft Excel, PowerPoint or Word tables and graphics, tableR supports the export in all of these formats.

likert: An R Package for Visualizing and Analyzing Likert-Based Items

Kimberly K. Speerschneider[1,2,⋆]
Jason M. Bryer[1,2]

1. University at Albany
2. Excelsior College
⋆Contact author: kimkspeer@gmail.com

Keywords: likert, questionnaire data, visualization, grammar of graphics

The Likert (Likert, 1932) item format has become the *defacto* standard in survey research. In the most common format of Likert-items, respondents rate their agreement with a statement from strongly disagree to strongly agree, usually with four to seven levels. Rensis Likert assumed that the distance between each response category are equal, and as such, analysis has typically treated the responses to Likert-items as continuous variables. However, this assumption often does not hold (see e.g. Wakita et al., 2012), although can often easily be verified with the use of visualizations. This talk introduces the **likert** package that provides a set of functions for analyzing Likert-items, visualizing results using the **ggplot2** (Wickham, 2009) package, and reporting results with the **xtable** (Dahl, 2012) package. Figure 1 represents one such graphic analyzing reading attitudes from the Programme of International Student Assessment (PISA; OECD, 2010)

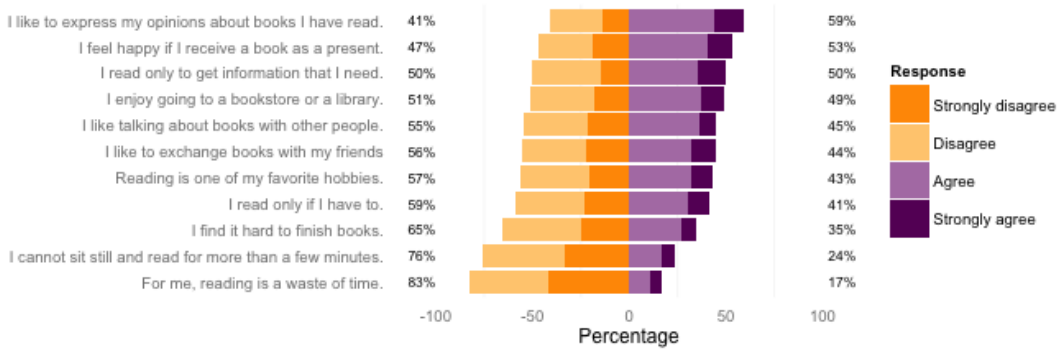


Figure 1: Attitudes Towards Reading

References

Dahl, D. B. (2012). *xtable: Export tables to LaTeX or HTML.* R package version 1.7-0.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology 142*, 1–55.

OECD (2010). Programme of international student assessment (pisa).

Wakita, T., N. Ueshima, and H. Noguchi (2012). Psychological distance between categories in the likert scale : Comparing different numbers of options. *Educational and Psychological 72*.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Design of graphics with lattice and `mosaic`

Richard M. Heiberger[1,⋆]

1. Temple University
⋆Contact author: rmh@temple.edu

Keywords: likert, lattice, mosaic

The **lattice** and **vcd** packages are two powerful paradigms for programming multi-dimensional graphical structures.

I recently completed two implementations of Diverging Stacked Bar Charts for Likert Scales and Other Applications (Robbins and Heiberger, 2011; Heiberger and Robbins, 2013). Both are included in the **HH** (Heiberger, 2013) package in *R* (R Development Core Team, 2013). One, `likert`, is based on `barchart` in **lattice** (Sarkar, 2012, 2008). The other, `likertMosaic`, is based on `mosaic` in **vcd** (Meyer et al., 2012, 2006; Zeileis et al., 2007).

Some features are more easily programmed in one setting than the other. Variable-width bars are natural in the `mosaic` setting and difficult in the `barchart` setting. Extensive secondary labeling seems easier in the **lattice** setting. Numeric axes are more natural in the **lattice** setting. Conditioning factors are possible in both with very different ways of thinking about them. Different sets of levels in the response factor and in conditioning factors are handled very differently by the two settings. The formula interface appears more powerful in the **lattice** setting than in the `mosaic` setting. Examples and solutions will be illustrated in both settings.

References

Heiberger, R. M. (2013). *HH: Statistical Analysis and Data Display: Heiberger and Holland*. R Foundation for Statistical Computing. R package version 2.3-36.

Heiberger, R. M. and N. B. Robbins (2013). Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software submitted*, 1–36.

Meyer, D., A. Zeileis, and K. Hornik (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software 17*(3), 1–48.

Meyer, D., A. Zeileis, and K. Hornik (2012). *vcd: Visualizing Categorical Data*. R Foundation for Statistical Computing. R package version 1.2-13.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robbins, N. B. and R. M. Heiberger (2011). Plotting likert and other rating scales. In *JSM Proceedings*, Section on Survey Research Methods, Alexandria, VA, pp. 1058–1066. American Statistical Association.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer. ISBN 978-0-387-75968-5.

Sarkar, D. (2012). *Lattice Graphics*. R Foundation for Statistical Computing. R package version 0.20-6.

Zeileis, A., D. Meyer, and K. Hornik (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics 16*(3), 507–525.

Bosco-R: Empowering Users to Create On-Demand R Resources

Fraser, Dan[1], Weitzel, Derek[2] , Gore Brooklin[3] , Ahronovitz, Miha[4]

1. Argonne Lab
2. University of Nebraska
3 Computer Science, University of Wisconsin, Madison
4. Bosco Team, University of Chicago

*Contact author: mihaa@uchicago.edu

Keywords: on-demand, clusters, R, resources, multiple jobs, not limited, high throughput, HTC

The Open Science Grid is developing a capability that enables **R** users to dynamically create computational resources for running **R** jobs. This capability is primarily directed at users who have multiple R jobs that need to run simultaneously, and are limited by the resources currently available to them. All that is needed is an account on a local campus cluster and a Mac or Linux based desktop for submitting jobs. The campus cluster does Not need to have **R** pre installed. Bosco-R does that for you. The basic idea is as follows: Users download Bosco R and install it on their (Mac or Linux) desktop system. By typing a simple command users can "add" clusters to their desktop by entering their account and password on the cluster. Then users submit jobs to Bosco, and Bosco transparently installs **R** and sends the jobs out to run on the available resource clusters. After the jobs have completed, Bosco-R brings the data back to the users' desktop for analysis.

In this paper we describe these capabilities and how to get users started creating on-demand R resources and running multiple simultaneous R jobs..

References

Bosco Web Site: *http://bosco.opensciencegrid.org/*

Open Science Grid provided the funding*: http://bosco.opensciencegrid.org/*

Derek Weitzel Blog: *http://derekweitzel.blogspot.com/*

Miha Ahronovitz Blog : How Bosco simplifies your life
http://my-inner-voice.blogspot.com/2013/02/how-bosco-simplifies-your-life-on-amazon.html

Practical computer experiments in *R*

Yann Richet[1,*] , Miguel Munoz Zuniga[1]

1. Institut de Radioprotection et de sûreté Nucléaire, France
*Contact author: yann.richet@irsn.fr

Keywords: HPC, computer experiments, experimental design, computing

The “computer experiments” field is an active topic of research relying on many statistical tools well supported by *R*. Nevertheless, a practical implementation faces technical issues to launch the underlying mechanistic simulations, which are mostly not related to R, often high CPU consuming and sometimes unavailable outside a remote cluster. We give an overview of some possibilities to perform such computer experiments locally from R, focusing at ergonomic issues the user will encounter.

Considering computer experiments as a regular task of experimental design implies a repetitive by-hand modeling, which is surely the most versatile and simplest way to start from. Nevertheless, even ignoring the hardship for the user, and the induced risk of error, it may not be so straightforward when the design is iterative like for an optimization algorithm.

A first computing-level solution is to wrap the simulator into a *R* function, such that the call for its evaluation is transparent from the perspective of the design algorithm. This point may hold several degrees of complexity when considering “local” vs. “remote”, “one by one” vs. “vectorized” or “short” vs. “long” simulations. An upstream approach concerns the architecture of the algorithm code, which requires to split each simulations iteration batch, so the user controls the sequence loop of data exchange between R and the simulator.

We will discuss both solution advantages and limitations, and conclude with some practical examples from nuclear safety applications. A supplementary application case is to be presented in a related session through the new GPC sensitivity analysis package.

References

Richet (2010). Promethee project – [R] user interface, http://promethee.irsn.fr/doku.php?id=user:r

The ReDICE Consortium (2011), http://www.redice-project.org

Roustant, Ginsbourger, Deville (2012), DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization, *Journal of Statistical Software*, 51.

Richet, Caplin, Crevel, Ginsbourger, Picheny (2012) Using the Efficient Global Optimization Algorithm to assist Nuclear Criticality Safety Assessment, *Nuclear Science and Engineering*, in press.

Symbiosis – Column Stores and *R* Statistics

Hannes Mühleisen[1,*], **Thomas Lumley**[2]

1. Database Architectures Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
2. Department of Statistics, University of Auckland, Auckland, New Zealand
*Contact author: hannes@cwi.nl

Keywords: Databases, Column Store, Virtual Data Object

A typical work flow for data analysis in *R* consists of the following steps: First load the raw data from file, then select and transform raw data into a form suitable for statistics, and then apply a statistical algorithm and visualization. However, the amount of data that can be analyzed using this process is limited by the amount of memory on the system on which *R* is run, which are typically desktop computers. A logical next step to mend this problem is to store the raw data in a relational database system. The standard process is now modified by not loading the raw data into *R*, but instead to load it into a database. Then, one can "outsource" the selection of data relevant to the analysis as well as basic calculations and aggregations to a highly optimized database system.

R's database interface (**DBI**) provides a generic way of communicating with a relational database. Packages such as **RPostgreSQL** implement a specific driver for a particular database. However, not all relational databases are equally well suited to support statistical calculations. Transformation procedures and simple calculations make recommending a relational database optimized for "On-line analytical processing" (OLAP) rather obvious. Furthermore, *R*'s calculations on statistical observations are typically performed column-wise. Hence, only a fraction of columns are actually processed at a given time. These factors together suggest a column-oriented database design. MonetDB, an open-source column-oriented database system, implements this design. We have created the **MonetDB.R** package, which implements a native **DBI** driver to connect *R* with MonetDB.

However, in order to tell the database which data is to be transferred to *R*, a user still is required to write queries in the standardized Structured Query Language (SQL), which breaks work flows and increases training requirements. We went one step further and implemented a *virtual data object*. This `monet.frame` object is designed to behave like a regular *R* `data.frame`, but does not actually load data from MonetDB unless absolutely required. For example, consider the following interaction: `mean(subset(mf,c1 > 42)$c2)`. We select a subset of the codemf object based on a filter condition on the `c1` column. Then, we average of the `c2` column. However, in this case the `mf` variable points to an instance of our virtual data object backed by a MonetDB table `t1`. Our implementation automatically generates and executes a SQL query: `SELECT AVG(c2) FROM t1 WHERE (c1>42);`. Instead of loading the potentially large table, we only transfer a single scalar value. Also, through the columnar storage layout of MonetDB, only the files that contain the data for columns `c1` and `c2` actually have to be accessed.

Our approach has two major advantages: Users are not exposed to SQL queries at all, and only data relevant to the analysis are loaded int *R*, which results in huge performance improvements. `monet.frame` is part of **MonetDB.R**, and we invite all those interested to take part in its evolution.

References

Idreos, S., F. Groffen, N. Nes, S. Manegold, K. S. Mullender, and M. L. Kersten (2012). MonetDB: Two decades of research in column-oriented database architectures. *IEEE Data Engineering Bulletin 35*(1), 40–45.

Mühleisen, H., T. Lumley, and A. Damico (2013). MonetDB.R. http://cran.r-project.org/web/packages/MonetDB.R/.

Memory Management in the TIBCO Enterprise Runtime for R (TERR)

Michael Sannella[1,*]

1. TIBCO Software Inc.
*Contact author: msannell@tibco.com

Keywords: R, memory management, performance

Most people using *R* don't need (or want) to know about its internal architecture: How it represents data objects, allocates memory, and frees unused objects. Sometimes, though, the internal details of memory management make a difference when trying to write efficient *R* code. Experience shows that many *R* performance problems are best viewed as memory problems.

TIBCO has recently released the *TIBCO Enterprise Runtime for R* (*TERR*), a new *R*-compatible engine. Our team had a unique opportunity to redesign and rebuild the internal data representation and memory management facilities from scratch, and we attempted to address long-standing problems with the internal architecture of *R* and related systems (*S* and *S+*). This talk will describe design decisions we made developing *TERR*, and discuss how they affect time and memory efficiency when executing *R* code. Having another *R*-compatible engine to compare with *R* presents a new perspective that may inform *R* engine development in the future.

TiddlyWikiR: an R package for dynamic report writing.

David Montaner[1⋆], **Francisco García-García**[1,2]

1. Biostatistics Department. Institute of Computational Medicine. Centro de Investigación Príncipe Felipe. Valencia. Spain.
2. Spanish Institute of Bioinformatics.
⋆Contact author: dmontaner@cipf.es

Keywords: Dynamic report, wiki.

TiddlyWiki is a single page wiki application. It is built in a unique HTML file which includes CSS and JavaScript besides the document content. As any other wiki system, users may add, modify, or delete content using a web browser. Being a wiki, it has the advantage over plain HTML pages of the simplified markup language and the easiness of edition. But unlike most other wiki applications, TiddlyWiki does not need any installation; it does not even need being hosted in a web server. The single file that constitutes the application is downloaded and kept locally while the edition is ongoing. It can be used as a local document or it may be finally uploaded to a server and made accessible via Internet as any other HTML file.

TiddlyWiki content organization relies on chunks of information called *tiddlers*. Tiddler can be set to be display in the document when it is first opened, or it can be accessed through the appropriated links when necessary. This feature makes TiddlyWiki optimal for writing small statistical reports: a main document can be display linearly by default while complementary information as for instance the explanation of the statistical glossary can be kept in the background and accessed just when needed by the reader.

Being a single file a TiddlyWiki document can be straightforward used as a template for such statistical reports. First the wiki system will allow for the quick edition of the text and for the specification of the document lay out. Then, as TiddlyWiki is ultimately a text file, automatic routines may insert additional information into the report as for instance tables of descriptive statistics, results form hypothesis testing or links to plots that will be displayed within the document.

TiddlyWikiR is an *R* package for writing dynamic reports using TiddlyWiki as a template. It implements S4 classes to organize and handle statistical results within R, and functions to insert those results into the wiki file.

TiddlyWiki is published under an open source license which makes it very suitable to the *R* users community.

References:

http://tiddlywiki.com

http://bioinfo.cipf.es/dmontaner/tiddlywikir

Synthesis of Research Findings Using R

Randall E. Schumacker, Ph.D.
The University of Alabama
rschumacker@ua.edu

Keywords: Meta-Analysis, Research, Statistics

An individual researcher spends time and resources to conduct a single study. The research is often published in a peer reviewed professional journal. Over time, many research studies will have been published on a specific topic (Cooper, 1998). The ability to synthesize research findings across a number of studies on a single topic provides an overall assessment of the cumulative effect size (Cooper, Hedges, & Valentine, 2009). An *R* script was written to provide two different methods of synthesizing research findings: use of *p*-values or the use of a test statistic, *df*, and *p*-values in a study (Schumacker & Tomek, 2013). Results show an *r* effect size, a *d* effect size, and an unbiased *d* effect size.

Bax, Yu, Ikeda, and Moons (2007) provided a recent comparison of many meta-analysis software program features (*Comprehensive Meta-analysis (CMA), MetAnalysis, MetaWin, MIX, RevMan*, and *WEasyMA*). Today, many of these features and capabilities exist in R packages: **ma_meta-analysis** (Del Re & Hoyt, 2010). The 5 packages are: **compute.es**, **MAc**, **MAc GUI**, **MAd**, and **MAd GUI**. *R* has made conducting meta-analysis user friendly by offering a GUI interface with pull down menus for Mac and Windows computers. The reference manuals, *ggplot2* function to display results, and the **R2wd** package to export formatted tables in Word are useful tools when using these packages . The R packages will be further discussed.

References

Bax, L., Yu, Ly-Mee, Ikeda, N., & Moons, K. GM (2007). A *systematic comparison of software dedicated to meta-analysis of causal studies*. *BMC Medical Research Methodology*, *7* (40), 1471-2288. (http://www.biomedcentral.com/1471-2288/7/40)

Cooper, H. (1998). *Synthesizing Research* (3rd Edition). *A Guide for Literature Reviews*. Applied Social Research Methods Series, Volume 2. Sage: Thousand Oaks, CA.

Cooper, H., Hedges, L.V., Valentine, J.C. (2009). *The Handbook of Research Synthesis and Meta-Analysis* (2nd Edition). Russell Sage Foundation: NY, NY.

Del Re, A.C. & Hoyt W.T. (2010). *MA* Packages*, http://rwiki.sciviews.org/doku.php?id=packages:cran:ma_meta-analysis

Schumacker, R.E. & Tomek, S. (2013). *Understanding Statistics Using R*. Springer-Verlag: NY, NY.

compreGroups updated: version 2.0

Isaac Subirana[3,1,4*] **Héctor Sanz**[2,1], **Joan Vila**[1,3]

1. Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, IMIM, Barcelona
2. Barcelona Centre for International Health Research (CRESIB, Hospital Clínic-Universitat de Barcelona), Barcelona, Spain
3. CIBER Epidemiology and Public Health (CIBERESP), Spain
4. Statistics Department, University of Barcelona, Spain

*Contact author: isubirana@imim.es

Keywords: Software Design, Bivariate Table, LaTeX, Descriptive Analysis

The package **compareGroups** is an utility tool available on CRAN designed to build tables containing descriptions of several variables stratified by groups that can be displayed in a clear, easy to read format on the *R* console, or included in a LaTeX report, or exported to CSV or HTML file. This package was first presented at the useR!2010 conference [1], and published a beta-version on CRAN. We presented some new innovations at useR!2011 [2], and finally released version 1.0. Thanks to users' feedback, this package has been debugged and much improved, and the current version (2.0) incorporates several important changes and innovations to improve functionality, versatility and reach:

- **compareGroups website:** New website with package description, vignette, examples, news, feedback questionnaires, subscribers list, forum, etc.
- **Web user interface:** Dynamic web interface, allowing users to run **compareGroups** on a remote *R* session, without knowledge of *R*-scripting or installation.
- **Reporting:** New function to output a full report of results.
- **Extended vignette:** The vignette has been extended in line with new functionality and data examples.
- **Genetic analysis:** New capability for SNP data, including reading and coding raw data using functionality from the **SNPassoc** package [3], and quality control processes, and summary outputs.
- **New data set:** Data from a longitudinal RCT with >7,000 individuals. The PREDIMED study [4].
- **Output formatting:** Improved customisation of output tables, including number of decimals, categorisation, and other character formatting.

Table 1: Univariate associations of risk factors from PREDIMED study

	Males			Females		
	[ALL] N=3165	HR	p.ratio	[ALL] N=4282	HR	p.ratio
Demographic and intervention group variables:						
Intervention group:						
Control diet	987 (31.2%)	Ref.	Ref.	1463 (34.2%)	Ref.	Ref.
Mediterranean diet with VOO	1050 (33.2%)	0.72 [0.50;1.02]	0.065	1493 (34.9%)	0.69 [0.45;1.06]	0.091
Mediterranean diet with nuts	1128 (35.6%)	0.60 [0.41;0.87]	0.007	1326 (31.0%)	0.80 [0.51;1.25]	0.334
Age	66.0 (6.55)	1.06 [1.04;1.09]	<0.001	67.7 (5.83)	1.10 [1.06;1.13]	<0.001
Risk factors:						
Smoking:						
Never	835 (26.4%)	Ref.	Ref.	3729 (87.1%)	Ref.	Ref.
Current	804 (25.4%)	1.47 [0.93;2.32]	0.099	243 (5.67%)	1.79 [0.96;3.34]	0.066
Former	1526 (48.2%)	1.75 [1.16;2.62]	0.007	310 (7.24%)	0.36 [0.12;1.15]	0.084
Type 2 diabetes	1708 (54.0%)	1.69 [1.23;2.33]	0.001	1906 (44.5%)	2.10 [1.44;3.07]	<0.001
Waist circumference	103 (9.60)	1.01 [0.99;1.03]	0.252	98.4 (10.8)	1.01 [0.99;1.03]	0.174

References

[1] Héctor Sanz, Isaac Subirana, and Joan-Salvador Vila. Bivariate analyses. In *useR! 2010, The R User Conference (National Institute of Standards and Technology, Gaithersburg, Maryland, US)*, July 2010.

[2] Héctor Sanz, Isaac Subirana, and Joan-Salvador Vila. comparegroups package, updated and improved. In *useR! 2011, The R User Conference (University of Warwick, Coventry, UK)*, August 2011.

[3] Juan R González, Lluís Armengol, Elisabet Guinó, Xavier Solé, , and Víctor Moreno. *SNPassoc: SNPs-based whole genome association studies*, 2012. URL http://CRAN.R-project.org/package=SNPassoc. R package version 1.8-5.

[4] R. Estruch, E. Ros, J. Salas-Salvadó, MI. Covas, D Pharm, D. Corella, F. Arós, E. Gómez-Gracia, V. Ruiz-Gutiérrez, M. Fiol, J. Lapetra, RM. Lamuela-Raventos, L. Serra-Majem, X. Pintó, J. Basora, MA. Muñoz, JV. Sorlí, JA. Martínez, MA. Martínez-González, and the PREDIMED Study Investigators. Primary prevention of cardiovascular disease with a mediterranean diet. *N Engl J Med*, Feb 2013.

BayesVarSel. An *R* package for Bayesian Variable Selection

Anabel Forte[1,3,⋆], **Gonzalo García-Donato**[2]

1. Universitat Jaume I
2. Universidad de Castilla La-mancha
3. Bayestats

⋆Contact author: forte@uji.es

Keywords: Bayes Factors, Model Space, Objective Bayes, Robust Priors.

BayesVarSel provides tools for solving Variable selection problems in the context of linear models and from an Objective Bayesian point of view.

BayesVarSel provides a user-friendly interface combining priors which are proved to give good theoretical results with computational advances to assess variable selection. In particular the prior distribution for the parametric space can be choosen among Liang et al. (2008); Zellner and Siow (1980, 1984); Zellner (1986); Bayarri et al. (2012) with the robust prior of Bayarri et al. (2012) being the default one. This prior have many well studied features for model selection and at the same time is available in closed form which allows for much faster computations.

BayesVarSel allows the calculations to be performed either exactly –`Bvs`(sequential) or `PBvs` (paralel computation)– or heuristically –`GibbsBvs`– using a Gibbs sampling algorithm studied in García-Donato and Martínez-Beneito (2013).

Most of the code is written in *C* and depends on:*R* ($\geq$ 2.15.0), **snow** and , **MASS**

References

Bayarri, M., J. Berger, A. Forte, and G. García-Donato (2012). Criteria for bayesian model choice with application to variable selection. *Annals of Statistics 40*(3), 1550–1577.

García-Donato, G. and M. Martínez-Beneito (2013). On sampling strategies in bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association In press*.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008). Mixtures of g-priors for bayesian variable selection. *Journal of the American Statistical Association 103*, 410–423.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g- prior distributions. In *In Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, pp. 389–399.

Zellner, A. and A. Siow (1980). Posterior Odds Ratio for Selected Regression Hypotheses. In *In Bayesian Statistics 1*, pp. 585–603.

Zellner, A. and A. Siow (1984). *Basic Issues in Econometrics*.

1

Bayesian learning of model parameters given matrix-valued information, using a new matrix-variate Gaussian Process

Dalia Chakrabarty, University of Leicester, University of Warwick, U.K.
Sourabh Bhattacharya, Indian Statistical Institute, India

Keywords: Bayesian non-parametric learning, Gaussian Process, Matrix-variate distributions, Transformation-based MCMC.

Abstract: The learning of model parameters given available data, is a problem that we often encounter, in all disciplines of science. Let all the scalar model parameters that we wish to learn be collated to define the model parameter vector $\boldsymbol{S} \in \mathcal{S} \subseteq \mathbb{R}^d$. In this work, we seek to learn $\boldsymbol{S}$ given the data that is in the form of a matrix. The information is then expressed as an unknown, matrix-variate function of the model parameter vector $\boldsymbol{S}$ and this unknown function is modelled using a high-dimensional Gaussian Process. The methodology is developed to deal with the availability of both training as well as test (or measured) data.

In particular, our interest lies in predicting the unknown value $\boldsymbol{s}^{(new)}$ of $\boldsymbol{S}$ that supports the measured data vector $\mathbf{v}^{(test)}$, given the training data matrix $\mathcal{D}_s$ that is generated at chosen values $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star$, of $\boldsymbol{S}$; we define the chosen design set as $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star\}$. The unknown relationship between the available information and $\boldsymbol{S}$ is trained on $\mathcal{D}_s$, which is generated at this design set. In fact, in our treatment the information is considered as a vector of corresponding dimensions, so that this function is modelled as a vector-variate Gaussian Process. This leads to the likelihood $[\mathcal{D}_{aug}|\boldsymbol{\Psi}]$ being rendered matrix-normal, where $\boldsymbol{\Psi}$ are relevant parameters of the Gaussian process and the training data matrix augmented by the measured data vector $\mathbf{v}^{(test)}$ is $\mathcal{D}_{aug}^T = (\mathcal{D}_s^T \vdots (\mathbf{v}^{(test)})^T)$. The mean and covariance matrices of the likelihood function are suggested by the structure of the Gaussian Process in question. Using this likelihood, we then construct the joint posterior of $\boldsymbol{s}^{(new)}$ and selected process parameters (such as the smoothness parameters) given the augmented data. Thus, the method allows for the learning of certain process parameters, in addition to the unknown model parameter vector value that supports the real data $\mathbf{v}^{(test)}$. Inference is performed using Transformation-based MCMC.

An application of this method is made to learn feature parameters of the Milky Way, using measured and simulated data of velocity vectors of stars that live in the vicinity of the Sun. Learning of the Galactic parameters with the real data is shown to produce a similar result to a comparator method that requires a much larger data set, in order to accomplish estimation.

`FluDetWeb`: an interactive web-based system for the early detection of the onset of influenza epidemics

David Conesa[1,*], **Antonio López-Quílez** [1], **Miguel Ángel Martínez-Beneito**[2], **Francisco Verdejo**[1]

1. Departament d'Estadística i Investigació Operativa, Universitat de València, 46100 Burjassot (Valencia), Spain
2. Centro Superior de Investigación en Salud Pública, 46020 Valencia, Spain
*Contact author: david.v.conesa@uv.es

Keywords: Bayesian hierarchical models, Hidden Markov models, Influenza, Sentinel Networks, Web-based surveillance systems.

The early identification of influenza outbreaks has became a priority in public health practice. A large variety of statistical algorithms for the automated monitoring of influenza surveillance have been proposed, but most of them require not only a lot of computational effort but also operation of sometimes not-so-friendly software.

In this paper, we introduce `fludetweb`, an implementation of a prospective influenza surveillance methodology based on a client-server architecture with a thin (web-based) client application design. Users can introduce and edit their own data consisting of a series of weekly influenza incidence rates. The system returns the probability of being in an epidemic phase (via e-mail if desired). When the probability is greater than 0.5, it also returns the probability of an increase in the incidence rate during the following week. The system also provides two complementary graphs. This system has been implemented using statistical free-software (`R` and WinBUGS), a web server environment for Java code (*Tomcat*) and a software module created by us (*Rdp*) responsible for managing internal tasks; the software package *MySQL* has been used to construct the database management system. The implementation is available on-line from: `http://www.geeitema.org/meviepi/fludetweb/`.

The ease of use of `fludetweb` and its on-line availability can make it a valuable tool for public health practitioners who want to obtain information about the probability that their system is in an epidemic phase. Moreover, the architecture described can also be useful for developers of systems based on computationally intensive methods.

References

M.A. Martínez-Beneito, D. Conesa, A. López-Quílez, A. López-Maside (2008). Bayesian Markov switching models for the early detection of influenza epidemics. Statistics in Medicine, 27(22):4455–4468.

D. Conesa, A. López-Quílez, M.A. Martínez-Beneito, M.T. Miralles, F. Verdejo (2009). BMC Medical Informatics and Decision Making, vol. 9, number 36.

Looking for (and finding!) hidden additivity in complete block designs with the `hiddenf` package.

Jason A. Osborne[1,⋆], **Christopher Franck**[2]

1. North Carolina State University, Department of Statistics
2. Virginia Polytechnic Institute and State University, Department of Statistics
⋆Contact author: jaosborn@ncsu.edu

Keywords: Interaction, Additivity, Tukey, Block Designs

A new test of additivity is presented for two-factor experiments with one observation per factor level combination. The hypothesis of additivity, or absence of interaction, is frequently of interest in complete block designs and many tests have been developed in the literature to assess its plausibility in such settings. The new test is based on a search for a 'hidden additivity' structure, where levels of one factors may be grouped such that within a group the effects of both factors are additive. Membership of levels within groups is treated as a latent variable. The computation of the test statistic and pvalue, along with a plotting procedure, are included in a new package under development entitled `hiddenf`.

The `hiddenf` package builds upon another recently developed package called `additivityTests`, which computes test statistics and critical values (but not p-values) from several well-known tests and also a new modification of Tukey's procedure [3]. `hiddenf` can call the functions in `additivityTests` and use them them to compute p-values. It also expands the collection by adding several new tests, including those by [2] and [1]. Lastly, it has plotting functionality that enables the user not only to test for interaction, but to characterize it with graphical assistance.

The methodology and use of the package are illustrated using several datasets taken from the statistics literature on interaction. Additionally, a study of copy number variation in dogs with lymphoma is presented which serves as a rich source of two-factor data with one observation per factor level combination.

References

[1] Franck, C. (2010). *Latent Group-Based Interaction Effects in Unreplicated Factorial Experiments*. Ph. D. thesis, North Carolina State University.

[2] Kharrati-Kopaei, M. and S. Sadooghi-Alvandi (2007). A new method for testing interaction in unreplicated two-way analysis of variance. *Communications in Statistics- Theory and Methods 36*, 2787–2803.

[3] Simecek, P. and M. Simeckova (2013). Modification of tukey's additivity test. *Journal of Statistical Planning and Inference 143*, 197–201.

A ggplot2 builder for Eclipse/StatET and Architect

Willem Ligtenberg[1,⋆], **Stephan Wahlbrink**[2]

1. OpenAnalytics BVBA
2. WalWare
⋆Contact author: willem.ligtenberg@openanalytics.eu

Keywords: ggplot2, GUI, Architect, Eclipse, StatET

The **ggplot2** package (Wickham, 2009) implements a highly popular and feature-rich graphics system for *R*. Since it uses a layered grammar of graphics (Wickham, 2010), it lends itself very well for access via a graphical user interface and until recently two such user interfaces were available to R users. The first interface is the web-based ggplot2 builder of Jeroen Ooms (Ooms, 2010). The second interface is the Plot Builder that is part of the Deducer GUI of Ian Fellows (Fellows, 2012). The first interface, however, is a standalone tool that lives separately from the other tools a user may be using to develop R code. The second interface, on the other hand, is part of more general R GUI effort which (legitimately) neglects general tools for editing R code.

In this presentation a new **ggplot2** builder will be presented that integrates smoothly with the Eclipse based R IDE offered by the StatET plugins (Walware, 2013) and Architect (OpenAnalytics, 2013). The plot builder offers similar functionality to the other interfaces, but is strongly influenced by the Eclipse philosophy in terms of user interface design. We will demonstrate the features of the plot builder in detail and show how the integration with the IDE allows to maintain the reproducibility of the analysis even if graphical user interfaces come into play.

References

Fellows, I. (2012). Deducer: A data analysis gui for R. *Journal of Statistical Software˜49*(8), 1–15.

Ooms, J. (2010). ggplot2 web application: A web interface for the r package ggplot2. http://rweb.stat.ucla.edu/ggplot2.

OpenAnalytics (2013). Architect: a state of the art interface with state of the art statistical technology. http://www.openanalytics.eu/architect.

Walware (2013). StatET: Eclipse plug-ins for R. http://www.walware.de/goto/statet.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics˜19*(1), 3–28.

Visualizing Multivariate Contrasts

Jason Waddell[1,*]

1. OpenAnalytics BVBA
*Contact author: jason.waddell@openanalytics.eu

Keywords: Visualization, Graphics, Contrasts, Mixed-Effects Models

Multivariate analyses often require large numbers of contrasts to be computed between levels of an explanatory variable across many timepoints or experimental conditions. The visualization literature, however, does not pay much attention to such very specific graphs and data analysts currently rely on sub-optimal extensions of basic plots.

We present two new clean and intuitive methods for visualizing model contrasts, along with several examples of visualizing mixed-effects models in practice.

metaplot: Flexible Specification for Forest Plots

Edna Lu[1], **Paul Murrell**[1], **David Scott**[1,*],

1. University of Auckland
*Contact author: d.scott@auckland.ac.nz

Keywords: Meta-analysis, forest plot

Current implementations of forest plots in R packages allow users to produce and modify forest plots from meta-analysis outputs within the same package. However, for such a complex plot, flexibility is often restricted to the set of options the functions provide.

Metaplot allows the construction of extremely flexible forest plots and implements the grid naming scheme to allow for specific customisations. User preferences in drawing forest plots are implemented without needing laborious specifications of options or resorting to modifying the function code. In addition, the design also ensures that minimal effort is required from the user when no customisations are required

GaRGoyLe: A map composer using *GRASS*, *R*, *GMT* and *LaTeX*

Francisco Alonso-Sarra[1,⋆]

1. Universidad de Murcia. Instituto del Agua y del Medio Ambiente
⋆Contact author: alonsarp@um.es

Keywords: Mapping, GRASS, GMT, LaTeX

GaRGoyLe is a mapping system based on *GRASS*, *R*, *GMT* and *LaTeX* (all of them open source software). It includes a set of*R* functions which main objective is to create maps in PDF format from *GRASS* raster or vector layers. However it can be adapted to work with other programs.

Most Geographic Information Systems (GIS) include tools to design and compose maps. However, these tools have two main drawbacks. They are mainly based on button sequences, so they are difficult to program, and they allow little creativity by imposing a quite rigid layout. On the contrary, *GaRGoyLe* functional approach allows the user to conceive the mapping process as a program; and, using *LaTeX* to design the layout, the creativity limit is on the user's expertise with this program. However, the basic use of *GaRGoyLe* is quite simple and does not require *LaTeX* skills. One of the objectives of the functions included is to be easy to use. This aim is achieved by using default values for most of their parameters. In this sense, *GaRGoyLe* tries to be more like GRASS modules or R functions than like GMT modules.

GaRGoyLe has two main functions:

- *GMT_map*. This function uses *GMT* to generate a PDF file from a set of *GRASS* display commands. This file contains the main map image, with the scale and paper size determined by the user, including a legend, a grid and annotations, if so decided by the user.
- *latex_map*. This function generates and compiles a *LaTeX* file that includes the PDF file generated by the previous function. It allows the user to add other cartographic elements as a title, a scale bar, text boxes, other images, etc. Most of these elements are defined as chunks of LaTeXcode.

Other functions creates *GMT* legend or label files from *GRASS* files and creates or modifies *GRASS* colour files .

GaRGoyLe uses *R*, *GRASS*, *GMT* and *LaTeX*, these programs should be installed in the system; however, only R is really critical. You can use any PNG file with a map instead of a set of GRASS commands to create the base map. *LaTeX* is needed by `latex_map` but you can obtain simple maps just with `GMT_map`. The R package **spgrass6** is needed to access the GRASS geometric information (`GRASS region`) and the package **classInt** is needed by some of the functions that deals with color palettes. Several *LaTeX* packages are used to set the map dimensions (**a0poster**, **sciposter**), to put elements in the map (**textpos**) and to introduce some graphics (**tikz**). Some of the functions that interact with *GRASS* vector layers were conceived assuming that the *PostgreSQL* driver is being used to store tables. If it is not the case they may not work properly.

This map programming capacity becomes specially useful when lots of maps with a common layout have to be created in thesis or consultancy projects. In addition, the flexibility introduced by *LaTeX* allows to fulfil difficult design specifications on map layouts or to create conference posters inserting several maps. In summary, **GaRGoyLe** is an *R* package to give *GRASS* users the power and flexibility of *GMT*, *R* and *LaTeX* languages to create maps.

Asymmetric Volatility Transmission in Airline Related Companies in Stock Markets

Eui-Kyung Lee

Department of Business Administration, Daejin University, Republic of Korea

Keywords: Granger cause, GARCH, Stock Return

This study examines the interrelationship of the stock return of airline related companies using GARCH model. The airline related companies include three fields of companies; airline companies, airport companies, and travel companies. The results of this research show three findings. The first is that airline stock return affects airport stock return. But the same is not true in reverse. The second is that airline stock return affects travel stock return significantly. The same is not true in reverse here either. The third is that no meaningful relationship could be found between airport and travel companies. The findings of this research can be useful in making policies on the airline related industry and this study contributed to the diversity of airline industry researches.

References

[1] Eun, Cheol S. and Sangdal Shim, "International Transmission of Stock Market Movement,"*Journal of Financial and Quantitative Analysis,* 24(2), (1989), 241-256.

[2] Becker, Kent G., Joseph E. Finnerty and Manoj Gupta, "The Intertemporal Relation between the U. S. and Japanese Stock Markets," *Journal of Finance,* 45(4), (1990), 1297-1306.

[3] Koutmos, Gregory and G. Geoffrey Booth, "Asymmetric Volatility Transmission in International Stock Markets," *Journal of International Money and Finance,* 14(6), (1995), 747-762.

A R tool to teach descriptive statistics

Encarnación Álvarez[1], **Antonio Arcos**[2,⋆], **Juan F. Muñoz**[2], **María M. Rueda**[2]

1. Centro Universitario de la Defensa. Universidad Politécnica de Cartagena
2. University of Granada
⋆Contact author: arcos@ugr.es

Keywords: Individual exercises, Self-assessed exercises, teaching tool.

Descriptive statistics is an important discipline inside statistics that describes the main features of a collection of data. This topic is common in many subjects from many areas of any University. This paper presents a teaching tool based on *R* that provide to the students self-assessed and individual exercises about descriptive statistics. Students can use this tool to check a set of exercises, which are based on the three personal numbers A, B, C, which consist of the three last number of the National Identity Document (DNI) of students. In this way the set of exercises is individual. In addition, the teaching tool provide a code to the student when the set of exercises is properly solved. This personal code is used by the teacher to check the whole set of exercises, which implies that the proposed teaching tool is also an instrument to save an important amount of time to the teacher when checking sets of exercises. As commented, this teaching tool has been created using the statistical software *R*, and it has been used in different groups of the University of Granada with satisfactory results.

Using *R* to estimate parameters from multiple frames

Antonio Arcos[1,⋆], M. del Mar Rueda[1], Giovanna Ranalli[2] and David Molina[1]

1. Department of Statistics and Operational Research. University of Granada, Spain
2. Department of Economics, Finance and Statistics, Università degli Studi di Perugia, Italy
⋆ Contact author: arcos@ugr.es

Keywords: Dual frames, Hartley estimator, single frame approach

The use of more than one single list of population units is important because one of the common practical problems in conducting sample surveys is that frames that can be used for selecting the samples are generally incomplete or out of date, so that if a sample is drawn from them, it may not be representative of the entire population and the survey is therefore affected by possible serious bias. On the other hand, if containing the entire population, it may be very expensive to sample from and updating a list is a difficult and very expensive operation that, even if it has become easier due to the recent advances in managing databases, it always requires an important and expensive effort in data collection.

Recently, multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. Much attention has been devoted to the introduction of different ways of combining estimates coming from the different frames.

However, most statistical commercial software does not incorporate estimation procedures for multiple frames. In this paper we present various functions in *R* that implement some important estimators like those proposed by Hartley (1962) and by Fuller and Burmeister (1972), the Pseudo Maximum Likelihood estimator proposed by Skinner and Rao (1996) and the recently proposed Pseudo Empirical Likelihood estimator (Rao and Wu, 2010). Estimators proposed under the so called single-frame approach (Bankier, 1986) are implemented as well.

References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 1074–1079.

Fuller, W.A. and Burmeister, L.F., (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of social science section of The American Statistical Asociation*,

Hartley, H.O.(1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

Rao, JNK and Wu, C., (2010) Pseudo–Empirical Likelihood Inference for Multiple Frame Surveys. *Journal of the American Statistical Association*, pp. 1494–1503

Skinner, C.J. and Rao, JNK., (1996) Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, pp. 349–356.

Calibration in Complex Survey using *R*

A. Arcos*, M. Rueda, D. Molina

Department of Statistics and Operational Research. University of Granada. Spain
*Contact author: arcos@ugr.es

Keywords: Calibration, Complex Surveys.

There are a growing for demand information on all kinds of topics. Such information can be collected and compiled in a survey. This can be an expensive and time-consuming process. Fortunately, rapid developments in computer technology have made it possible to conduct more complex surveys.

The first step is the design of the survey. The second step in the process is data collection. Finally, a clean file is obtained which is ready for analysis. And in this last step, some weighting technique are applied. The characteristics of the correction weights should be such that the weighted estimator has better properties than the HorvitzThompson estimator. The auxiliary information can also be used to compute weighting adjustments.

In the context of Sample Surveys, Deville and Sarndal (1992) and Deville et al. (1993) have created a general framework for weighting of which linear and multiplicative weighting are special cases. The starting point is that adjusted weights have to satisfy two conditions: a) the correction weights have to be as close as possible to 1 and b) the weighted sample distribution of the auxiliary variables has to match the population distribution. The first condition sees to it that resulting estimators are unbiased, or almost unbiased, and the second condition guarantees that the weighted sample is representative with respect to the auxiliary variables used.

Deville and Sarndal (1992) show as the minimization problem can be solved by using the method of Lagrange. By choosing the proper distance function, both linear and multiplicative weighting can be obtained as special cases of this general approach. Deville et al. (1993) showed that estimators based on weights computed within their framework have asymptotically the same properties. This means that for large samples, estimators based on both weighting techniques will behave approximately the same. However, although the estimators behave in the same way, the individual weights computed by means of linear or multiplicative weighting may differ substantially. And the computational problems increase as sample size increases.

In this work we review some *R* functions from several packages for this purpose and perform various comparisons.

References

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association 88*, 1013–1020.

R/STATISTICA Interface

Peyman Eshghi[1,*]

1. GlaxoSmithKline
*Contact author: peyman.5.eshghi@gsk.com

Keywords: STATISTICA, RDCOM, REXCEL

R is a powerful, free, multi platform, and highly extendable statistical programming environment, where its usage is rapidly expanding through the statistical community in many specialised areas. Due to its extremely flexible nature, it is quite unrealistic to expect a comfortable graphical user interface within R. This can sometimes put off potential users in applied sciences and industry, from embarking on using *R*.

In recent years, some statistical packages, such as *Excel, SAS, and STATISTICA* have facilitated integration with R, allowing users to utilise the power of R within the user-friendly environment of a commercial applications. This has also allowed developers to combine the graphical user interface of these packages with the strong statistical analysis' capacity of R.

Non-expert users often prefer the comfort of a commercial package. *STATISTICA*, in particular is a point and click statistical package that allows macro development using a specific brand of *Visual Basic*. In this poster, I will demonstrate the integration of sophisticated techniques offered by *R* with the user interface of *STATISTICA*.

As a member of a statistical computing team in GlaxoSmithKline, I have written a number of user-friendly *STATISTICA* macros for scientists, where the bulk of the analysis is performed in R.

References

Thomas Baier & Erich Neuwirth, Powerful data analysis from inside your favorite application. http://rcom.univie.ac.at/

STATISTICA v8, StatSoft Ltd. STATISTICA, http://www.statsoft.co.uk/

AMOEBA+ with R

Guillermo Valles Castellano[1,⋆], **Antonio López-Quílez**[1], **Raquel Pérez-Vicente**[2], **Juan Merlo**[2]

1. Statistics and Operational Research Department, University of Valencia
2. Faculty of Medicine, Lund University (Malmö, Sweden)
⋆Contact author: guivacas@alumni.uv.es

Keywords: Spatial Clustering, Getis-Ord statistic, Grids

0.0.1 Introduction

The spatial analysis is very important in any academic field that needs the spatial structures to explain an outcome of interest. For this purpose we have established collaboration between the Unit for Social Epidemiology at the Faculty of Medicine of Lund University and the Statistics and Operational Research department at the University of Valencia. The collaboration is developed as a project called "A longitudinal multilevel analysis of socioeconomic disparities in cardiovascular diseases: questioning past evidence with new methodological approaches". To try to solve this problem we need an algorithm that could identify the irregular clusters present in the data. For this reason, we use A Multidirectional Ecotope-Based Algorithm (AMOEBA).

0.0.2 Origins of the AMOEBA

This procedure uses the Getis-Ord local statistic G_I^*. AMOEBA uses this statistic to test the null hypothesis of no association between the value found at a place and its neighbors within the designated region. The improvements of the algorithm decreased the computation time.

0.0.3 AMOEBA+

We do not just implement the constructive AMOEBA in R, we also improve it. This improvement rises in the resolution of the risk levels, i. e., from the original algorithm you can only obtain three levels of risk. However, this new improvement can return us more than the three original levels of risk. The improvement lies in re-apply the AMOEBA at different levels that are obtained. Note that AMOEBA and AMOEBA+ are implement to apply to a political structure or a grid over the region.

0.0.4 Application

We apply both methods in two examples. The first example is the Number of Holdings from the 2009 Agrarian Census. We focus in the whole number of holdings in Castilla-La Mancha. The second example represents the results of applying both algorithms, original and evolved, to the incidence of the respiratory disease in Skåne. We apply the algorithms to a grid with 64 cells. That are 64 cells per side so we have 4096 cells.

0.0.5 Conclusions

The original algorithm of AMOEBA have been tested in some publications. We want you to note that although three categories can be enough for little regions, it could not be enough for bigger regions. In our opinion, using only three categories is a poor approach to explain patterns in a huge region, so we developed the AMOEBA+, that provides the possiblity to get as many categories as needed.

Software developments for non-parametric ROC regression analysis

María Xosé Rodríguez - Álvarez[1], Javier Roca - Pardiñas[1,⋆], Carmen Cardarso - Suárez[2]

1. Department of Statistics and Operations Research. Universidade de Vigo, Spain
2. Unit of Biotatistics. Department of Statistics and Operations Research. Universidade de Santiago de Compostela, Spain
⋆Contact author: roca@uvigo.es

Keywords: Receiver operating characteristic curve, induced ROC regression methodology, direct ROC regression methodology, Non parametric regression model, *R* software

The ROC curve is a fundamental technique in the characterization of the accuracy of continuous diagnostic tests. Regression approaches of either the test results (Induced methodology) or the ROC curve itself (Direct methodology) have became the usual methods for the assessment of covariate effects on the ROC curve. Recently, new non-parametric estimators for the conditional ROC curve have been proposed, based on both induced and direct modelling (Rodríguez - Álvarez et al., 2011a,b). In this work we introduce a user-friendly software, called **npROCregression**, that implements these new approaches. The estimation procedure of both methodologies implies a high computational cost, especially as bootstrap methods are used for inference purposes. As a result, the programming language selected to implement these approaches was *Fortran*. However, to facilitate the use in practice, *R* (R Core Team, 2013) was chosen as the user interface program. The software offers numerical and graphical output for the estimated ROC curve, jointly with other summary measures of the accuracy, such as the area under the curve (AUC) or the generalized Youden index (YI). The software also provides the thresholds values based on the YI criterion and the criterion of equal sensitivity and specificity.

References

R Core Team (2013). R: A language and environment for statistical computing. http://www.R-project.org/. ISBN 3-900051-07-0.

Rodríguez - Álvarez, M. X., J. Roca - Pardiñas, and C. Cadarso - Suárez (2011a). A new fexible direct roc regression model - application to the detection of cardiovascular risk factors by anthropometric measures. *Computational Statistics and Data Analysis 55*, 3257–3270.

Rodríguez - Álvarez, M. X., J. Roca - Pardiñas, and C. Cadarso - Suárez (2011b). Roc curve and covariates: extending induced methodology to the non-parametric framework. *Statistics and Computing 21*, 483–499.

An *R*-package for Weighted Smooth Backfitting with structured models

Javier Roca - Pardiñas[1,⋆], **Stefan Sperlich**[2], **María Xosé Rodríguez - Álvarez**[1]

1. Department of Statistics and Operations Research. Universidade de Vigo, Spain
2. Department of Economics. Research Center for Statistics. Université de Genève, Switzerland
⋆Contact author: roca@uvigo.es

Keywords: generalized structured models, weighted smooth backfitting, generalized varying coefficients, generalized additive models, *R* software

An *R* (R Core Team, 2013) software package, called **wsbackfit**, is introduced that provides estimation procedures for a large class of regression models common in applied statistics. The regression models belong to the class of the so-called generalized structured models. They include for example generalized varying coefficient and generalized additive models, both optionally partial linear. The estimation procedure is based on smoothed backfitting which to our knowledge is the statistically most efficient existing procedure for this model class (Mammen et al., 1999; Roca - Pardiñas and Sperlich, 2010). Additional weights allow for both, the inclusion of different link functions and the efficient estimation under heteroscedasticity. Also, a cross validation bandwidth selector is provided.

References

Mammen, E., O. Linton, and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics 27*, 1443–1490.

R Core Team (2013). R: A language and environment for statistical computing. `http://www.R-project.org/`. ISBN 3-900051-07-0.

Roca - Pardiñas, J. and S. Sperlich (2010). Feasible estimation in generalized structured models. *Statistics and Computing 20*, 367–379.

Using R as continuous learning support in Sea Sciences Degree

Jose Zubcoff[1*], Francisco Gomariz-Castillo[1,2], Jose Vicente Guardiola[1] and Aitor Forcada[1]

1. Departamento de Ciencias del Mar y Biología Aplicada, Universidad de Alicante, 03080, Alicante, España
2. Instituto Euromediterráneo del Agua, Campus de Espinardo, s/n, 30001 Murcia, España
*Contact author: jose.zubcoff@ua.es mailto:

Keywords: bioinformatics, biostatistics, statistics learning, education research

This work presents how *R* is used in the development of skills within the Sea Sciences Degree at the University of Alicante. *R* is used as an aid in the gradual development of the student curriculum: from basics statistics in the first year, to complex experimental design, multivariate analysis or application to other fields such as geostatistics, in the latter years of the degree. Usually, the use of different software along a degree hinders the student's acquisition of knowledge, hence it is desirable to homogenize the use of tools, in order to focus efforts on learning concepts/methodologies. In this regard, we strongly recommend the use of *R* as a tool for support learning due to its great versatility.

The use of *R* along the course is justified by several advantages: *i)* it is free and open source software; *ii)* it is cross-platform, which facilitates its use by students; *iii)* it has great potential for data manipulation; and *iv)* the large and dynamic community that gives support to *R* and also the amount of documentation that is easily accessible. The supporting community has focused, not only on the implementation of new treatment techniques and data analysis, also on the development of packages designed to facilitate the access and consultation to some specific information, such as **rfishbase**, a client access to the specialized database FishBase, widely used in Marine Sciences.

We introduce *R* in the first year Statistics subject. *R* is used to initiate students into basic concepts of descriptive statistics and univariate and bivariate statistical inference (goodness of fit contrasts, contrasts normal populations, one-way ANOVA, contrasts for proportions, measures of association and linear regression) to allow the student to provide a basis for dealing with the resolution of problems in subsequent courses. In this course, students are also introduced into aspects of importing data or connection to databases using **rodbc**. Later, in the third year of the degree, within the subject "Statistics applied to marine resources", *R* and specifics packages are used to learn about experimental design, multifactorial ANOVA (using package **GAD**), and several multivariate analysis such as multiple linear regression, cluster, MDS, PCA, … (using packages **faraway**, **vegan**, **scatterplot3d**, **rgl** and **bpca**). Finally, as an optional subject, in "Geographic Information Systems and Remote Sensing" *R* is used for support both, GIS and remote sensing, due to its interoperability with other Open Source softwares like *GRASS and QGIS*. In this case, *R* is used for teaching geostatistics, through the wide variety of packages available for management and analysis of spatial data (**maptools**, **rgdal** or **spgrass6** to work with spatial information and **sp**, **gstat** or **automaps** for spatial interpolation and geostatitics). In this course, *R* is also used for the spatial distribution of species by glm models, and as a complementary tool for analysis by specific libraries to support several tasks, such as the analysis of spectral signatures or supervised and unsupervised classification in remote sensing.

As a result of the formative development, at the end of the degree, the student must have acquired among other competences: be able to apply mathematical and statistical knowledge to biology and the different application areas (resources, planning and management, modeling, etc.), understand and apply basic aspects of sampling and data processing, get the ability to recognize and solve different problems using these tools and, be able to communicate the results of the experiments. Furthermore, the use of *R* will be also a valuable skill acquired by students.

Variable selection algorithm implemented in FWDselect

Marta Sestelo[1⋆], Nora M. Villanueva[1], Javier Roca-Pardiñas[1]

1. Department of Statistics and Operation Research, University of Vigo
⋆Contact author: sestelo@uvigo.es

Keywords: variable selection, regression models, bootstrap, package

In a multivariate regression framework, the target response Y can depend on a set of p initial covariates $X_1, X_2, \ldots, X_p$ but in practical situations, one has to decide which covariates are "relevant" to describe this response. A question that tends to arise in regression models is determining the best subset or subsets of q ($q \leq p$) predictors which will establish the model or models with the best predictive capability. This problem is particularly important when p is high and/or when there are redundant predictors. As a general rule, an increase in the number of variables to be included in a model provides an "apparently" better fit of the observed data. However, these estimates are not always satisfactory: a) inclusion of irrelevant variables would increase the variance of the estimates, resulting to a partial loss of the predictive capability of the model; b) inclusion of many variables would mean that the model would be difficult to interpret. To solve this problem, we introduce a new forward stepwise-based selection procedure, which includes the selection of the best combination of q variables and the determination of the number of them to be included in the model. This methodology is implemented in an *R* package, **FWDselect**, an alternative to existing approaches, in the form of a simple method whereby *R* users can select the best model to be applied to different types of data (continuous, binary or Poisson response) in different contexts (linear models, generalized linear models or generalized additive models).

Panel time series methods in R

Giovanni Millo[1,⋆]

1. Generali Research and Development
⋆Contact author: mailto:giovanni_millo@generali.com

Keywords: Panel data, Time series, Unit roots, Cointegration, Econometrics

In the last decade, econometricians have devoted increasing attention first to the issues of nonstationarity and cointgration in pooled time series data (Phillips and Moon, 1999), then to the related problem of cross-sectional dependence induced by common factors (Pesaran, 2006). Among different approaches, Pesaran's Common Correlated Effects (CCE) augmentation has proved particularly promising both for consistent estimation and unit root testing under nonstationarity and cross-sectional strong dependence. An extension of package **plm** (Croissant and Millo, 2008) is described, comprising pooled and heterogeneous CCE estimators in both the pooled mean groups and the generalized fixed effects (CCEP) styles. The same software framework allows for cross-sectional-correlation robust, so-called "second generation" unit root tests as cross-sectionally augmented Dickey-Fuller (CADF) regressions (Pesaran, 2007). A practical demonstration of panel cointegration and error correction model estimation highlights the usefulness of the existing **plm** infrastructure for lagging and differencing in this new context.

References

Croissant, Y. and G. Millo (2008). Panel data econometrics in r: The plm package. *Journal of Statistical Software 27*(2), 1–43.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics 22*(2), 265–312.

Phillips, P. C. and H. R. Moon (1999). Linear regression limit theory for nonstationary panel data. *Econometrica 67*(5), 1057–1111.

Teaching introductory statistics to students in economics: a comparison between *R* and spreadsheet

Xosé M. Martínez[1,*]

1. Departament of Applied Economics II, University of A Coruña
*Contact author: xose.martinez@udc.es

Keywords: : introductory statistics, teaching, economics

One of the main disadvantages in using *R* in introductory statistics courses is the need to explain the code to people who only handle software with a graphical user interface. In addition, the limited time available for teaching how to use any kind of software is a problem that leads to the regular use of spreadsheets as support software in our different subjects rather than more suitable statistics programs

However, even on an elementary level, all statistical packages provide a much greater range of options and are better suited to the needs of our disciplines. Hence, we decided to give students a basic introduction to R and spreadsheets, and allow them to choose which of the two to use in practice.

The practical sessions begin with an introduction, consisting of half hour for *R* and another half hour for the spreadsheet. During this time we run through the essentials and provide templates with codes and manuals that the learner groups adapt to their needs.

In the case of *R*, we explain the basics: the Rstudio interface, how to work with scripts, and basic use of variables (objects), and then focus on statistical concepts and how to calculate with *R*.

The session poster gives examples of the strategy being followed and the results obtained and the students' perception of their experience.

TestR: R language test driven specification

Petr Maj*, Tomas Kalibera, Jan Vitek

Purdue University, West Lafayette, IN
*Contact author: peta.maj82@gmail.com

Keywords: Formal Specification, Testing, Regression, Compatibility Testing

Every computer language that seeks global adoption and a large user base requires a precise specification. Language references, such as the *ECMAScript* reference for *JavaScript*, are mostly used to fulfill this aim. They provide a concise and systematic description of the language and are easy to read and understand by the programmers. However, similar to other forms of documentation, references can lag behind the actual implementation. This problem is even more pronounced in rapidly evolving systems where the costs of keeping the references synchronized with the implementation are simply too high. *R* is no exception to this rule. While it provides an excellent help system and reference guide, these sometimes disagree with the implementation and lack the clarity and precision of a formal specification.

TestR attempts to provide another solution to the specification problem. It is a large collection of tests aiming to map all features and corner cases of the *R* language in a way similar to *Java* Technology Compatibility Kit. The tests are structured by language features and come with textual descriptions so that they can be easily used as a human-readable behavioral specification, complete with examples. The tests provide a regression suite for the *R* implementation and this validation will be essential to allow identifying incompatible language changes. The tests can be run with different *R* implementations to allow for compatibility testing. Comparison of these findings will be presented.

Small area data visualization using ggplot2 library

Carlos Pérez-González[1,⋆], Enrique González Dávila[1] and Jesús Alberto González Yanes[2]

1. Departamento de Estadística e Investigación Operativa, Universidad de La Laguna
⋆Contact author: cpgonzal@ull.es
2. Instituto Canario de Estadística, Gobierno de Canarias

Keywords: small area estimation; spanish labor force survey; k-means clustering; synthetic estimator; relative mean square error.

The estimation in small areas has been of general interest for a long time by the increasing demand of information about social and economic indicators of various characteristics of a population. In particular, the Spanish Labor Force Survey (EPA) is a quarterly study carried out by the National Statistical Institute of Spain for the Canary Islands. This survey provides information of activity, unemployed and inactive totals for sex level at autonomous community and province level.

In the context of the Labor Force Survey, different techniques and methods of small area estimation have been applied to obtain precise estimations of activity, unemployed and inactive totals at province for sex level in the different sub-divisions, called ”counties”, of the islands. However, in this case, the number of regions represents a problem if we want to draw in a graphic the estimators (means, mean squared errors, coefficients of variation, etc.) to compare them each other. In this work, we use the **ggplot2** library of *R* to achieve a useful and nicely visualization of these data at sexes and regions level.

References

González-Dávila, E. (2007). CANAREA 2007: Estimaciones en áreas pequeñas. Memoria del proyecto CANAREA 2007 realizado en colaboración con el Instituto Canario de Estadística (ISTAC).

Morales, D. et al. (2007). Estimación en áreas pequeñas con datos de la encuesta de población activa en Canarias. *Estadística Española*, 49, pp. 301-332.

Rao, J. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons.

R as a Data Operating System for the Cloud

Karim Chine[1,*]

1. Cloud Era Ltd
*Contact author: karim.chine@gmail.com

Keywords: Cloud Computing, EC2, Collaboration, Analytics-as-a-Service, e-Learning

With public cloud computing, a new era for Research and Higher Education begins. Scientists, educators and students can now work on advanced high capacity technological infrastructures without having to build them or to comply with rigid and limiting access protocols. Thanks to the cloud's pay-per-use and virtual machine models, they rent the resources and the software they need for the time they want, get the keys to full ownership, and work and share with little limitation. In addition, the centralized nature of the cloud and the users' ubiquitous access to its capabilities should make it straightforward for every user to share with others any reusable artifacts and to interact with them in real time: This is a new ecosystem for open science, open education and open innovation. What is missing is bridging software.

We propose such software (*Elastic-R*) to help data scientists, educators and students take full advantage of this new ecosystem. With Cloud Computing, infrastructures became programmable and the provisioning and control of virtual compute and storage resources became accessible via standard Web Services. *R* became the lingua franca of data analysis and one of the most widely used tools for programming with Data. *Elastic-R* merges the capabilities of *R* and those of public and private clouds to enable new ways of interacting with data, building applications and services and collaborating with considerable power and flexibility. The *Elastic-R* platform transforms *Amazon EC2* into a ubiquitous and scriptable collaborative environment for traceable and reproducible data analysis and computational research. It makes the acquisition, use and sharing of all the capabilities required for statistical computing, data mining and numerical simulation easier than ever: The cloud becomes a user friendly Google-Docs-like platform where all the artifacts of computing can be produced by any number of geographically distributed real-time collaborators and can be stored, published and reused.

References

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

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

[3] www.elastic-r.net

TPmsm: Estimation of the Transition Probabilities in 3-State Models

Artur Araújo [1], **Luís Meira-Machado**[1], **Javier Roca-Pardiñas**[2,⋆]

1. Department of Mathematics and Applications, University of Minho
2.Department of Statistics an O.R., University of Vigo
⋆Contact author: roca@uvigo.es

Keywords: Survival, Kaplan-Meier, Multi-state model, Illness-death model, Transition probabilities

One major goal in clinical applications of multi-state models is the estimation of transition probabilities. The usual nonparametric estimator of the transition matrix for non-homogeneous Markov processes is the Aalen-Johansen estimator (Aalen and Johansen,1978). However, two problems may arise from using this estimator: first, its standard error may be large in heavy censored scenarios; second, the estimator may be inconsistent if the process is non-Markov. Happily, there have been several recent contributions that account for these problems. In this work we consider the estimation of the transition probabilities, using **TPmsm** a software application for *R*. It describes the capabilities of the program for estimating these quantities using seven different approaches. In two of these approaches the transition probabilities are estimated conditionally on current or past covariate measures. The software is illustrated using data from two real data sets.

References

Aalen, O. and Johansen, S. (1978). An Empirical Transition Matrix for non Homogeneous Markov and Chains Based on Censored Observations. *Scandinavian Journal of Statistics 5*, 141–150

Climate Analysis Tools
An operational environment for climate products

Rebekka Posselt[1,*], Sophie Fukutome[1], Mark A. Liniger[1]

1. Federal Office of Meteorology and Climatology MeteoSwiss
*Contact author: rebekka.posselt@meteoswiss.ch

Field: Environmetrics
Area: Visualization /Graphics

Keywords: automatic production, know-how transfer, climate, database

MeteoSwiss has established an internal framework using *R* to implement the analysis of climate data in an operational context. It aims at coordinated development, automatic production, and easy maintenance. Climate researchers develop new methods and create new visualizations for monitoring the current climate and analyze its past. These developments are integrated in the common framework of the Climate Analysis Tools (CATs). The CATs are organized as a collection of *R* packages with mandatory documentation. To ease maintenance, functions are structured in three separate categories of modules: data retrieval from the climate observation database, statistical analysis, and graphical, tabular or data output. Three types of *R*-packages are differentiated: *R* packages officially available from CRAN, *R*-packages for developers containing complex statistical algorithms developed internally, *R*-packages for users containing the user interface, the data retrieval, handling und output generation.

This strict structure allows efficient translation of scientific findings into operational products. The final CATs are installed in an operational environment managed by the IT department. Production is controlled via over 80 cron-jobs and monitored by IT professionals without *R* or climate expertise. Some examples can be found on the official webpage of MeteoSwiss, such as the portals for Climate indicators (MeteoSwiss, 2012) and for Trends at Stations (MeteoSwiss, 2010). Currently, the CATs are used to generate around 80’000 graphical and tabular products on a regular basis in a fully automatic manner.

This contribution will present the strategy and structure underlying the CATs and will discuss the experience with this framework as a link between climate research, statistics, and operational production for a federal agency.

References

MeteoSwiss (2012). Meteoswiss Climate indicators – Browser
http://www.meteoswiss.admin.ch/web/en/climate/climate_today/climate_indicators/indicators_browser.html

MeteoSwiss (2010).MeteoSwiss Trends at Stations.
http://www.meteoswiss.admin.ch/web/en/climate/climate_today/trends_from_stations.html

seq2R: Detecting DNA compositional change points

Nora M. Villanueva[1⋆], **Marta Sestelo**[1], **Javier Roca-Pardiñas**[1]

1. Department of Statistics and Operation Research, University of Vigo
⋆Contact author: nmvillanueva@uvigo.es

Keywords: kernel, bootstrap, DNA, change points, derivatives

Identifying compositional change points in a statistical framework can be a challenging task. Numerous methodological approaches have been developed to analyse change points models, i.e. Bayesian estimation, maximum-likelihood estimation, least squares regression or nonparametric regression. Part of our philosophy is to make easier for others to use a new statistical methodology. With that in mind, we implement in a user-friendly and simply *R* package, **seq2R**, a methodology that identifies and locates compositional change points in DNA sequences by fitting regression models and their first derivatives. Since the estimation procedure of this methodology with large datasets implies a high computational cost, Fortran is used as the programming language. Our approach to assess the change points, is based on detecting the maximum or minimum of the first derivative of nonparametric regression models with binary response. To estimate the regression curve and its first derivative, we apply local linear kernel smoothers. Additionally, bandwidths are automatically selected using cross-validation techniques. Inference implies the construction of confidence intervals which can be obtained by bootstrap methods. The choice of the bandwidth and the usage of bootstrap resampling techniques may entail high computational cost. To considerably reduce this cost and render the operational procedures, we apply binning techniques.

NPRegfast: Inference methods in regression models including factor-by-curve interaction

Marta Sestelo[1*], **Nora M. Villanueva**[1], **Javier Roca-Pardiñas**[1]

1. Department of Statistics and Operation Research, University of Vigo
*Contact author: sestelo@uvigo.es

Keywords: factor-by-curve interactions, kernel, bootstrap, binning, testing derivatives

In general, regression analysis plays a fundamental role in statistics. The purpose of this technique is to evaluate the influence of some explanatory variable on the mean of the response. In certain circumstances, the relationship between a continuous covariate and the response can vary among subsets defined by levels of a categorical covariate, resulting in a regression model with factor-by-curve interaction. Our package, **NPRegfast**, introduces an estimation method for this type of models along with different techniques for drawing inferences about them. The package enables *R* users to compare the regression curves specific to each level, and even to compare their critical points (maxima, minima or inflection points) through the study of their derivatives. The main estimation procedure is based on local polynomial kernel smoothers. Inference with this package (confidence intervals and tests) is based on bootstrap resampling methods. Accordingly, binning acceleration techniques are also implemented to ensure that the package is computationally efficient.

Pharmaceutical market analysis with R

Anna Bednarczyk[1]*, Dorota Marszałek[1]

1. Kielce University of Technology, Poland

*Contact author: ania8746@o2.pl

Key words: pharmaceutical market, statistical data analysis, templates relevant reports

In the case of the pharmaceutical industry, the statistical analysis of the data is used at every stage of therapeutic agents: from the search for new active substances for drug development, clinical trials, release of new products to production and current and periodic monitoring of the process, sell and profit from the individual preparations.

Using the R software, the analysis of the issues listed below was performed:

i. impact and the role of changing conditions in the market economy on the pharmaceutical industry,
ii. distribution of pharmaceutical product sales in terms of therapeutic group,
iii. reason for the discrepancy in the sale of OTC medicines in different countries,
iv. trend increase in the price and quantity of medication,
v. medication-often chosen by patients.

After analysis in R we get the results in the form of tables and graphs. The program provides a flexible and convenient mechanisms for the management of such results, so that they can be placed in a folder or directly in the report, which can then be saved as a Word or pdf file.

Templates relevant reports were created on the basis of statistics carried out by operators in pharmaceutical companies, so that the analysis obtained contains commonly used standards.

References

Michael Bartholow (2011). Top 200 Drugs of 2010.
Aleksandra Baranowska-Skimina (2010).
http://www.egospodarka.pl/59769, market-products-OTC-2010-2012, 1,39,1. html OTC Market 2010-2012 (in Polish).
Interpharma ph, the Association of Swiss pharmaceutical companyPetersgraben 35, 4003 Basel(in Polish).
Malgorzata Michalik, Bogna Pilarczyk, Henryk Mruk, Strategic marketing in the pharmaceutical market, Issue 2, Wolters Kluwer Poland,sp. z o.o.(in Polish).

Standardisation on Statistics: ISO Standards and R Tools

Emilio L. Cano [1,4,*], Javier M. Moguerza[1,4], Iván Moya Alcón[2,4], Mariano Prieto Corcoba[3,4]

1. Universidad Rey Juan Carlos
2. Asociación Española de Normalización (AENOR)
3. ENUSA Industrias Avanzadas, S.A.
4. AENOR CTN66/SC3 Member
*Contact author: emilio.lopez@urjc.es

Keywords: Standardisation, ISO Standards, Regulatory Compliance, Quality Science

The process of developing and implementing technical standards, or standardisation, is usually unknown, and often even disregarded. Nontheless, in specific areas of activity, such as Electrical Engineering or Aeronautics, practitioners are more concerned as technical standars are mandatory in many cases. However, in crossing disciplines as Statistics and other science-related topics, we seldom care about standards, even though research projects funding schemes (e.g. FP7 and the approaching Horizon 2020) are more and more requiring the adoption and creation of standards. In this work, we outline the ISO Standards development process and their national implementation. We focus on statistics-related standards, with an outlook to the *R* domain: capabilities, opportunities and challenges.
The ISO Standards development is carried out by Technical Committees (TC) which deal with specific subjects. The ISO/TC69 (Applications of statistical methods) is in charge of the development of the standards related to our field. It is organised into six subcommittes. The authors are involved with the AENOR Standardisation Technical Committe AEN/CTN66/SC3 (*métodos estadísticos*), whose aim is to update and adopt in Spain International Standards stemmed from ISO/TC69.
The ISO/TC69 directly develops some of the standards that are not in charge of the subcommittes, for example ISO 28640 (Random variate generation methods). It defines several methods to generate random variates, wich can be compared with *R* Random Number Generation (RNG) methods and eventually certify that our method fulfills an international standard. Likewise TC69/SC1 (Terminology and symbols) develops ISO 3534 (Vocabulary and symbols), which has 4 parts, and can be used as a reference for reports and research, and to assess **base** R functionalities. Many *R* packages contain functions that are related to ISO standards and can be assessed (and eventually certified), for example: ISO 7870 series, Control Charts (**qcc**, **SixSigma**, **IQCC** packages); ISO 22514 series, Statistical methods in process management -- Capability and performance (**qualityTools**, **SixSigma** packages); ISO 3951 series, Sampling procedures for inspection by variables (**AcceptanceSampling** package); ISO 5725 series, Accuracy (trueness and precision) of measurement methods and results (**SixSigma** package); ISO 13053, Quantitative methods in process improvement -- Six Sigma (**SixSigma** package).
Undoubtedly, Data Scientists, and practitioners in general, can take advantage of the synergies between the adoption of international standards and the use of *R*. *R* programs can be easily verified (which is a requirement in many standards, e.g., ISO 9001) as it is Open Source. Thus, any company or organization can go beyond the traditional « Quality Certified » stamp and extend the use of international standards to their processes.

References

Asociación Española de Normalización - AENOR, http://www.aenor.es/ .

Emilio L. Cano, Javier M. Moguerza and Andres Redchuk (2012). *Six Sigma with R*. Springer, New York, http://www.sixsigmawithr.com/

ISO/TC69 Applications of statistical methods, http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=49742

Quantitative Text Analysis of readers' contributions on Japanese daily newspapers

Yasuto NAKANO[1,⋆]

1. School of Sociology, Kwansei Gakuin University, JAPAN
⋆Contact author: yasuto@soc-nakano.net

Keywords: text mining, correspondence analysis, readers contributions

0.0.1 Purpose

The purpose of this paper is to mine the big data of readers' contributions in Japanese daily newspapers. Methodologys used here are text mining and correspondence analysis.

0.0.2 Data

There are several daily newspapers distributed throughout Japanese nation wide. Each of these papers is circulated for round 10 millions. The biggest two papers cover almost 40 % of Japanese population. In such newspapers, there is a page for readers' contributions, placing about 10 articles contributed by readers'. Occasionally there is a special topic offered by the editor, but in general, the topic of contributions depends on contributers' concerns. We have set up cumulative data of this page for 20 years. One contribution includes not only text of the articel but also its contributer's name, age, occupation and address. We extract these contributer's attributes from the article text, and make them related to it. Analyzing cumulative contributions data tells us opinions of certain groups of Japanese. It tells us transitions and stability of opinions. Furthermore, this data tells us relations between topics and contributers' attributes.

To analyze Japanse text data, we have to conduct morphological analysis on the data. We use **RMeCab** package, which utilize a morphological engine 'MeCab' from *R*. **RMeCab** provides us a Document Term Matrix(DTM). In case of 'ASAHI', there are 50 thousands contributions. We morpholize about 70 thounsands different words. Therefore our data of 'ASAHI' consists of a matrix of 50 thousands article rows and 70 thousands words culumns. A matrix of occupation and words, 3 thousands occupations rows and 70 thousands words columns, is also available. These matrix can be visualized with the package **ca** and **igraph**.

0.0.3 Analysis

For example, we choose topic of 'war'. Amongst all articles, 8 % of 'ASAHI' articles includes 'war'. Frequencies of war-related articles are changed by month. August is a month of 'war'. 'Teachers' tend to mention on 'war' more frequently than other occupations. The DTM tells us what kind of words are closely related to 'war'. It reveals how contributers have descrived on 'war'. Words of 'Misery' and 'peace' have kept colose relation for 20 years. 'Responsibility' had close relation to 'war' 20 years ago, but it is reraly appeared in war artilce nowadays. Recently words of 'father' and 'mother' are conspicuous. Story of 'war' has been changed from contributers' own story to their parents' story. 'WWII' has been a main topic of war articles. But a word 'terror' appears a new topic in 00's.

This analysis is a example of textmining on our dataset with *R*.

Analysis of data from student surveys at Kielce University of Technology using R Commander and R Data Miner

Ewelina Nowak [1] , Monika Zuchowicz[1*]

1. Kielce University of Technology, Poland
*Contact author: m-zuchowicz@wp.pl

Keywords: survey data analysis, data visualization, **Rcmdr** package, **rattle** package

In Poland, at colleges and universities, there are nearly 2 million students. This is a group distinguished by its individual needs, not only for the material realm, but also immaterial one.

For several years, students of the Kielce University of Technology attending the course of statistics, at the beginning of the first semester fill up a simple questionnaire. Analysis of the data obtained from the survey provides important information for the students, as well as for the lecturer. The acquired data are also used to illustrate statistical methods presented during the course of statistics.

The students answer to several questions concerning among others expectations regarding the Polish economic situation in the next two years, opinions on how was changed their families financial situation over the past two years or self-assessment of their financial status. Students also inform about their approximate spending on cigarettes and alcohol, mobile phones, tourism and recreation. The questionnaire provides also information on the students gender, ethnicity, type of school completed, place of residence during the study.

Data obtained from the surveys that were conducted in the last five years have been subjected to in-depth analysis with the tools from the **Rcmdr** and **rattle** packages. These tools were used to perform descriptive statistical analysis, to visualize data, and to discover the relationships among variables. The study also looked at how the views and opinions of the students changed over time. The tools available from the Graphical User Interfaces, R Commander and R Data Miner, proved to be very efficient in discovering knowledge from the survey data.

References

J.H. Maindonald (2008). Using R for Data Analysis and Graphics – An Introduction, URL http://cran.r-project.org/doc/contrib/usingR.pdf.

John Fox (2005). The R Commander: A Basic-Statistics Graphical User Interface to R, *Journal of Statistical Software*, URL http://www.jstatsoft.org/v14/i09/paper

Graham J. Williams (2009). Rattle: A Data Mining GUI for R, URL http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf

Statistical analysis with R of an effect of the air entrainment and the cement type on fresh mortar properties

Julia Marczewska[1,*] , Wojciech Piasta[1]

1. Kielce University of Technology, Faculty of Civil Engineering and Architecture, Poland
*Contact author: jmarczewska@tu.kielce.pl

Keywords: energy savings, ecological modeling, design of experiments, **Rcmdr**, **rattle**.

Cement is an essential material for building and civil engineering. Unfortunately, its production is very energy-intensive. The greatest loss of energy in the cement industry is associated with the production of cement clinker, which is the main component (over 90%) of Portland cement CEM I. Poland, as a member of the EU is committed to reduce consumption of energy and CO_2 emissions. Both of these goals can be achieved through the use of additives in the production of clinker and cement.

Some products that are created in other industries as waste (e.g. blast furnace slag, silica fly ash) are usually used as additives. Cement additives eliminate large amount of waste from the environment and thereby reduce the amount of waste sent to landfill. The use of waste in the process of cement production decreases the consumption of energy for extraction and output of natural resources, but also reduces the share of clinker in cement.

In order to use efficiently cement additives we need to perform experiments to find optimal technology of cement production, as cement is mainly responsible for the quality of concrete. The influence of the type of cement and the air entrainment on properties of the fresh mortar was established by using five types of cements with different amounts of air-entraining admixture.

For statistical analysis of data obtained during the tests we used the packages **Rcmdr** and **rattle**. Regression analysis was performed to illustrate the relationships between the amount of air in fresh mortar and the amount of introduced impurities. Analysis of variance was used to demonstrate the effect of the type of cement and aeration on the diameter propagation and air content in fresh mortar mix. Statistical and graphical tools that are available from R Commander and R Data Miner, proved to be very useful in the analysis of the collected data.

Title

gxTools: Multiple approaches integrated in automated transcriptome analysis

Authors

Wolfgang Raffelsberger, Laetitia Poidevin, Hélène Polveche, Raymond Ripp,and Olivier Poch

Abstract

Transcription profiling has become a widely used technique in biomedical research and automated work-flows have emerged for routine data analysis. Over a wide range of different projects we have leaned that the choice of early data-treatment procedures -like normalization - has indeed a very strong impact on the results obtained by the various pipelines used, and furthermore, that a single treatment pipeline optimal for all biological questions seems not to exist. In the context of GxDb (http://http://gx.igbmc.fr/, a web based microarray database), we have set up a hierarchy of S4-class based objects to automatically generate, store and access multiple upload-protocols in parallel. In consequence this allows facilitated view and comparison of results by different workflows and to customize automated transcriptome analysis. Furthermore, our R-library gxTools is complemented by numerous tools for QC, diagnostic plots and/or custom follow-up of results.

A cloud infrastructure for *R* reports

Gergely Daróczi[1,2,3,⋆], **Aleksandar Blagotić**[4,5,⋆⋆]

1. Assistant lecturer at Pázmány Péter Catholic University, Hungary
2. PhD student at Corvinus University of Budapest, Hungary
3. Founder at Easystats Ltd, United Kingdom
4. Psychology MSc student at University of Niš, Serbia
5. Web and R developer at Easystats Ltd, United Kingdom

Contact authors: ⋆ daroczig@rapporter.net and ⋆⋆ alex@rapporter.net

Keywords: report, cloud, web application, server, security

The poster shows an annotated but mainly visual and rather technical overview of the infrastructure used at `rapporter.net` to generate reproducible statistical reports in a web application building on the power of *R* among other open-source projects.

Beside the replicated data stores (NoSQL databases and network drives), the problems of data conversion and the distributed back-end of *R* workers, the poster also concentrates on security issues like providing a *Rails* front-end, filtering user contributed *R* commands with **sandboxR** and evaluating all expressions in a **RAppArmor**-enforced temporary environment inside of **rApache** returning JSON – in short: presenting a successfull flowchart of open-source technologies after facing and resolving a variety of problems while creating *R*-driven web applications.

References

10gen (2013). MongoDB, `http://www.mongodb.org/`

Aleksandar Blagotić and Gergely Daróczi (2013). **rapport**: a report templating system. `http://cran.r-project.org/package=rapport`

Apache Software Foundation (2013). CouchDB, `http://couchdb.apache.org/`

Gergely Daróczi (2013). The **sandboxR** package: Filtering "malicious" Calls in *R*, `https://github.com/Rapporter/sandboxR`

Gergely Daróczi (2013). **pander**: An R Pandoc Writer, `http://cran.r-project.org/package=pander`

Jeffrey Horner (2013). **rApache**: Web application development with R and Apache, `http://www.rapache.net/`

Jeroen Ooms. The **RAppArmor** Package: Enforcing Security Policies in *R* Using Dynamic Sandboxing on Linux, JSS. In press. `https://github.com/jeroenooms/r-security`

John MacFarlane (2013). Pandoc, `http://johnmacfarlane.net/pandoc/`.

Rails Core Team (2013). Ruby on Rails, `http://rubyonrails.org`.

Red Hat, Inc. (2013). GlusterFS, `http://www.gluster.org/`

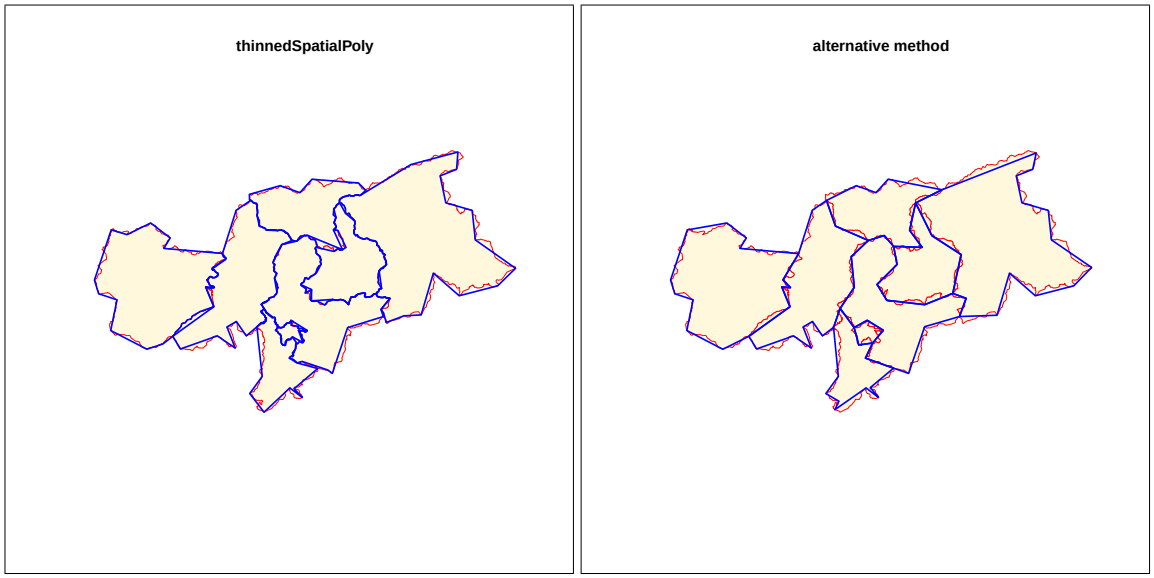
On thinning spatial polygons

Kurt Ranalter*

*Contact author: k.ranalter@gmail.com

Keywords: spatial polygons, thinning, neighbouring polygons, shared boundaries, slivers

We present a case study that deals with the thinning of spatial polygons. This operation is particularly useful when certain details can safely be omitted: for instance, when colouring a map it might well be sufficient to use simplified boundaries for administrative regions, i.e. boundaries that are just some approximation of the official ones. One advantage is that the reduced number of points helps to reduce the size of the graphical output. However, thinning spatial polygons is far from being trivial: the shared boundaries of neighbouring polygons might be treated differently and thus the resulting map may contain slivers. We propose a method that is similar in spirit to the `thinnedSpatialPoly` function of the **maptools** package. The main idea of our approach is to decompose each spatial polygon in such a way that it consists of various segments of its boundary, some shared some not. Each segment is simplified as it stands and then the information obtained during the decomposition is used to reconstruct the spatial polygons object.



Statistical analysis in R of environmental and traffic noise in Kielce

Krzysztof Maciejewski[1*] , Monika Stępień[1]

1. Kielce University of Technology, Poland
*Contact author: kmaciejewski@tu.kielce.pl

Keywords: noise, pollution, traffic, health, disorders,

Environmental noise pollution has grown to be a major problem in modern world. It relates to ambient sound levels that are perceived as uncomfortable and are caused by traffic, construction, industrial and recreational activities. In the long term, noise pollution may induce serious health problems such as hearing, sleep and mental disorders. Many government and international bodies have aimed their attention on identifying and monitoring local noise issues. The European Union puts a great emphasis on developing strategies to reduce the number of people affected by the noise.

The R environment is used for statistical analysis of noise levels acquired by an automated station located at the Sandomierska street in Kielce. Sandomierska while running through a heavily urbanized and residential districts, remains a major transport corridor for the eastern part of the city. The station is equipped with microphones (for noise measurements), traffic radars (for denoting speed, direction, position and type of vehicles) and is capable of gathering weather information. Such a broad spectrum of data allows to conduct extensive research regarding the influence of different factors on the recorded noise levels. The investigations include relationships between the noise and the traffic density, it's structure, weather conditions, day of the week and time of the day.

References

EU (2002). Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise. *Official Journal of the European Communities.*

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. *Springer*.

Lee, P., J., Shim, M., H., Jeon, J., Y. (2010). Effects of different noise combinations on sleep, as assessed by general questionnaire. *Applied Acoustics 71*, 870-875.

The report describes the application of graph theory in dosimetry.

Several of classical problems in graph theory in a reformulated form are applicable to the field of radiation safety.

We can represent a radiation situation on some territory in a form of undirected graph. Vertices of the graph form a regular grid, and edges connect neighboring vertices. Weight of each edge is equal to the dose, received by a person that has passed along this edge with some, fixed for the entire graph speed.

The report considers some problems of graph theory in their dosimetry interpretations.

The Shortest Path Problem answers the question how to go thru the radiation hazard area from point A to point B, and receive the lowest dose.

The traveling Salesman Problem determines order of passage the control points to obtain the minimum dose. For example points of planned determination of radiation situation.

Euler path problem helps to plan transport network bypass, so as to pass on every road once (where possible). This task is useful for determining the radiation situation using automatic dosimeter with georeference. In this case, the vertices of the graph are the crossroads, and the edges are the roads.

Route Inspection problem helps to plan transport network bypass, if there is no Euler path.

The problem of the shortest connection grid helps to connect some fixed number of dosimeters to one network in optimal way.

The problem of the critical path helps to choose the best option of a range of works, if the order of work is due by technological map of the enterprise.

R language is ideal for such tasks. The maximum useful packages are package "igraph", used to work with graphs; package "sp", allows to make georeferencing, and packages "raster", "rasterVis" and "rgl" for visualization.

Data mining with Rattle

Milena Nowek[1*], Justyna Jarmuda[1]**

1. Kielce University of Technology, al. Tysiaclecia Panstwa Polskiego 7, 25-314 Kielce, Poland
*Contact author: milena.nowek@gmail.com
**Contact author: juska88@wp.pl

Keywords: Rattle package, data mining, useful and clear information, calcium-silicate bricks

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives an summarizing it into useful information. Data mining software, such as package **Rattle**, is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationship identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [1]. Data mining combines concept, tools, and algorithms from machine learning and statistics for the analysis of very large datasets, so as to gain insights, understanding and actionable knowledge [2].

R is ideally suited to the many challenging tasks associated with data mining. **Rattle** (the *R* Analytical Tool To Learn Easily) is a graphical data mining application written in and providing a pathway into *R*. It has been developed specifically to ease the transition from basic data mining to sophisticated data analyses using a powerful statistical language. **Rattle** brings together a multitude of *R* packages that are essential for the data miner but often not easy for the novice to use. An understanding of *R* is not required in order to get started with **Rattle** – this will gradually grow as we add sophistication to our data mining [2], and this is what will be shown in our poster presentation – the way of exploring the analysis tools offered by **Rattle**, from basic to more advanced. We will try to present the most relevant statistical studies associated with marketing research and engineering datasets, both the essential and more complex. Our analyses are focused on modified calcium-silicate bricks. The main purpose is to present how a big number of information from huge databases can be combined, reduced and shown in the form of graphs and diagrams, and in what specific case it might be helpful. Simply saying – how to obtain the most desirable, useful and clearly expressed information from a large table of results of conducted research, using package **Rattle**.

References

Bill Palace (1996). Data Mining, http://www.anderson.ucla.edu.

Graham J. Williams (2009). Rattle: A Data Mining GUI for R, *The R Journal*, 45-55

intRegGOF: Modelling with the aid of Integrated Regression Goodness of Fit tests.

Jorge Luis Ojeda Cabrera[1,⋆]

1. Dep. Mtodos Estadsticos, U. de Zaragoza
⋆Contact author: jojeda@unizar.es

Keywords: Goodness of Fit, Integrated Regression, Modeling data, Cumulative Residuals, Marked Empirical Process.

Integrated Regression Goodness of Fit tests were introduced in Stute (1997) with the aim to detect the discrepancies between the actual regression model that data follows and any possible alternative. These tests are based on the fact that cumulative residuals along the values of the covariates characterizes the stochastic behaviour of the discrepancies between null and alternative hypothesis related to these regression models. In this way, these tests enable us to detect the lack of fit of a given model. As is pointed in previous reference the asymptotic distribution of functionals of this cumulative processes is not amenable and appropriate resampling methods in the regression framework are required to perform the tests.

R Package **intRegGOF** implements not only the main tools to perform Goodness of Fit tests based on Integrated Regression but also some utilities that allows modeling in a similar fashion to what is available in *R* for the class of models `lm`. In its present status, the package handles models developed using `lm`, `glm` y `nlm` tools for both unbiased and biased observations.

In this work we present the package, its main features and how are they implemented. The use of the *R* utilities to handle with several different models and computing on the language to develop the comparison between models is also addressed. Some examples with real and simulated data are also discussed to explain how the package works in both, the unbiased and selection–biased frameworks.

References

J. L. Ojeda, W. González-Manteiga, J. A. C. Testing regression models with selectionbiased data. *Submited for publication*.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist. 25*(2), 613–641.

Stute, W., W. González-Manteiga, and M. Presedo Quindimil (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc. 93*(441), 141–149.

W. Gonzlez-Manteiga, R. C. An updated review of goodness-of-fit tests for regression models. *TEST*, To be published.

An R script to model monthly climatic variables with GLM to be used in hydrological modelling

Francisco Gomariz-Castillo[1], **Francisco Alonso-Sarría**[2,⋆]

1. Instituto Euromediterráneo del Agua, Campus de Espinardo, s/n, 30001 Murcia, España; Dept. de Ciencias del Mar y Biología Aplicada, U. de Alicante, 03080, Alicante, España
2. Instituto del Agua y Medio Ambiente, Universidad de Murcia. Edificio D, Campus de Espinardo, s/n, 30001 Murcia, España
⋆Contact author: fjgomariz@um.es

Keywords: Spatial climatic models, GIS, Hydrological models, Generalised Lineal Models, Geostatistics

Despite the recent development of sophisticated interpolation techniques, systematic generation of climatic layers still uses traditional, local interpolation methods as the Thiessen polygons (also called Nearest Neighbour, NN) or the Inverse Distance Weighted (IDW). However, the impact of incorrect estimations of areal precipitation, precipitation variability, or even temperature on hydrological models is not negligible.

This work presents a methodology to model five climatic variables in an intensive and automatic way, from selection and correction of the series to spatial distribution modelling for each month and year. The variables analysed were average temperature, maximum absolute and average temperature, minimum absolute and average temperature and total precipitation. The time span of the study covers the period 1970-2002 with monthly temporal resolution.

All the process is organised around an *R* script to estimate the spatial distribution of the five climatic variables. Other programs used were *GRASS*, as a Geographic Information System and *PostgreSQL*, as a Database Management System. All the analysis was run on a shared memory linux supercomputer.

The R script calls different functions and uses several packages: **climatol** for previous homogenisation of climatological series, **RPostgreSQL** for access interface to data, **bestglm**, **lmtest** and **ipred** for select best GLM model and test GLM models, **Rmpi** for OpenMP implementation, **lattice** for powerful data visualisation, **xtable** for performance report results in latex, **spgrass6** for GRASS-R interface, **MBA**, **gstat**, **maps**, and **automap** for spatial and interpolation functions.

The method is based on a systematic application of Generalized linear models (GLM), using as predictors spatial location, elevation, slope, profile and tangential curvature, first order partial derivate E-W and N-S slope and second order partial derivates, sea distance and surface global radiation. To select the best model we used a simple exhaustive search algorithm and AIC provided in **bestglm** package. Finally, GML residuals were interpolated using different methods (inverse distance weighting, ordinary krigging and B-splines). The best interplation model was identified using leave-one-out cross-validation. These methods were also used as direct interpolators of the data to check in which conditions the GLM regression does not contribute to a better estimation.

Once the climatic variables were modelled, the statistical results were tested to select the best method of interpolation. Finally, a summary of the different methods used (test, goodness of fit, validation and selection of the best methods) was exported in latex format; and the spatial information was exported as inputs for a hydrological model implemented as a GRASS module.

From the results, we obtained general insights on the predictors that best explain the spatial pattern of the data, which families were more appropriate for each variable and which interpolation method obtained the minimum error.

Using R2wd package to automatize your reporting from *R* to *Microsoft Word* document – An application of automatic report for a survey in telecommunication

Céline Bugli[1*]

1. SMCS - Université catholique de Louvain, Belgium
*Contact author: celine.bugli@uclouvain.be

Keywords: R2wd, survey, automatic, report, reproducible research

In this talk we present examples of analysis of survey data integrated in the automatic report. We used the **R2wd** package to export automatically values, graphics and tables of an analysis of survey data into a *Microsoft Word* document with a pre-defined template. This package uses either the statconnDCOM server (via the **rcom** package) or the RDCOMClient to communicate with *MS-Word* via the COM interface. The survey was analyzing satisfaction of consumers about telecommunication. The analysis must be re-launch for each batch of received questionnaires over the year. This is why automation of the report is so important. Indeed, use of **R2wd** allows adapting your report automatically with template defined by the client. We will present examples of special commands provided in the **R2wd** package pushing output into a Word file.

Automation of spectroscopic data processing in routine tests of coals using *R*

Yuri Possokhov[1*]

1. Eastern Research & Development Institute of Coal Chemistry (VUKHIN), Yekaterinburg, Russia
*Contact author: possokhoff@gmail.com

Keywords: Spectroscopy of coals, Automation, Spectrometer software, Macro programming

Elaboration of automated high-accuracy tests of spectroscopic data in Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectrometry of coals is one of the most effective ways to increase the quality and the quantity of coal analysis in laboratories of coke and byproduct producers, cleaning plants, and mines [1, 2]. *R* with its *CRAN* gives new extensive opportunities in development of complex algorithms for processing spectroscopic data since *R* has the power revealed in its code clarity.

Modern spectrometer software includes conventional tools embedded for interactive spectrum processing. As for automation there are modules for macro programming by means of *VBA* or self-supporting development environment. However, the following shortcomings may be observed:

- complex algorithms of data processing (as those for processing DRIFT-spectra of coals) are always too difficult to develop as macro programs: either the programs embodied are large or the tools embedded are inadequate;

- once the macro program is embodied, it cannot be used with a FTIR-spectrometer of another brand.

In this regard it becomes obvious that using *R*-coding instead of macro programming completely solves the revealed shortcomings. Unfortunately, *R* environment cannot be embedded in the spectrometer software directly. Therefore, the elaboration of interface between *R* and the spectrometer software is the urgent problem nowadays.

So, we will describe the latest practical experience in automation of spectroscopic data processing in terms of Shimadzu IRsolution and *R* as applied to routine tests of coals.

References

[1] Possokhov Yu.M., Popov V.K., and Butakova V.I. (2010). Application of chemometric tools for coal calibration by DRIFT spectroscopy. In *Modern Methods of Data Analysis 2010, The Seventh Winter Symposium on Chemometrics, (Saint Petersburg, Russia)*, pp. 50–53.

[2] Bona M.T., Andrés J.M. (2008). Application of chemometric tools for coal classification and multivariate calibration by transmission and drift mid-infrared spectroscopy. *Analytica Chimica Acta 624, 68-78.*

A Web-based Application as a Dynamical Tool for Clinical Trial Researchers

Moisés Gómez-Mateu[1,⋆], **Guadalupe Gómez**[1]

1. Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Spain
⋆Contact author: moises.gomez.mateu@upc.edu

Keywords: Clinical trials, Composite endpoint, PluginR, Tiki Wiki software, Web-based application

We present a web-based application as a tool for investigators to help in the decision of the primary endpoint of a randomized clinical trial (RCT). The appropriate choice of the primary endpoint is a crucial issue, and often the decision is in terms of whether or not secondary endpoints have to be added in order to detect the desired effect of the treatment under investigation. Gómez and Lagakos (2013) developed a methodology to guide in the decision between using a composite endpoint or one of its components as the primary endpoint for efficacy. This method is based on the asymptotic relative efficiency and has already been programmed in *R*.

Our application is built by means of the software Tiki Wiki CMS/Goupware (Tightly Integrated Knowledge Infrastructure). This interface has been designed in such a way that scientists, without any knowledge of *R*, can interact with the application through HTML form pages. Users are asked to enter different inputs of information depending on the characteristics of their particular study. This information is recorded in trackers and processed in the server through *R* scripts. Dynamical results are shown by different scenarios with plots, together with reported recommendations to help in the choice between using a single or a composite endpoint as primary among a set of candidates.

References

Gómez, G. and Lagakos, S. (2013). Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in Medicine 32*, 719–738.

Sapir, R. (2010). Tiki Essentials. What every Smarty needs to know about Tiki Wiki CMS Groupware. *Under a Creative Commons Attribution-Share Alike 3.0 License.*

Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics 39*, 499–503.

Tiki wiki cms groupware. http://info.tiki.org/.

Analysis of load capacity of pipes with CIPP liners using R Rattle package

Kamil Mogielski[1,*] , Andrzej Kuliczkowski[1]

1. Kielce University of Technology, Faculty of Environmental Engineering, Geomatics and Power Engineering, Department of Water Supply and Sanitary Installations, Division of Water Supply and Sewage Systems, ul. Studencka 2, 25-314 Kielce, PL.
*Contact author: kamil.mogielski@gmail.com

Keywords: Survival analysis, 3D draw, Rattle, Cured-in-place pipe, Load capacity test.

Cured-in-place pipe (CIPP) is the most popular technology used to rehabilitate buried pipelines. In this technology, resin-soaked fabric is introduced into a damaged pipe and after that it is hardened.

Main purpose of the research was to check if there is any correlation between surface roughness of the pipe (depending on the material of it) and load capacity growth after pipe repair due to the sticky properties of the resin used for tube soaking. Concrete and vitrified clay pipes were tested.

Rattle package was used to show relationship between material of pipes and load capacity growth of both kinds of tubes. Survival analysis was used in order to explore forces applied to the samples at the moment of destruction. In presented analysis forces used to destroy the pipes made of concrete and vitrified clay were used as a parameter (Fig. 1a) instead of time. The graph shows that samples without liners that are made of concrete pipes have lower load capacity then analogic samples made of vitrified clay pipes. In samples with CIPP liners concrete pipes have bigger load capacity than clay pipes.

Rattle Package was also used in order to generate 3-D graphs illustrating relationship between three properties of samples: pipe material, internal dimension of pipe and thickness of epoxy liner. Sample 3-D graph is presented on Figure 1b. It shows that installation and thickness of epoxy liner have big influence to load capacity level. It is also clear that load capacity of concrete samples increase much more than in clay pipes although load capacities of pipes without liners were lower in the case of concrete pipes. Those sample analyses are only a part of research results. Other ones include decision trees, multi box plot graphs and more detailed survival analyses (for groups of samples with different nominal diameters, liner thicknesses and pipe materials).

Both presented analyses confirm hypothesis that material of renewed pipe should not be ignored during epoxy liner designing. It is necessary because load capacity of concrete pipes increase much more after liner installation than in vitrified clay pipes.

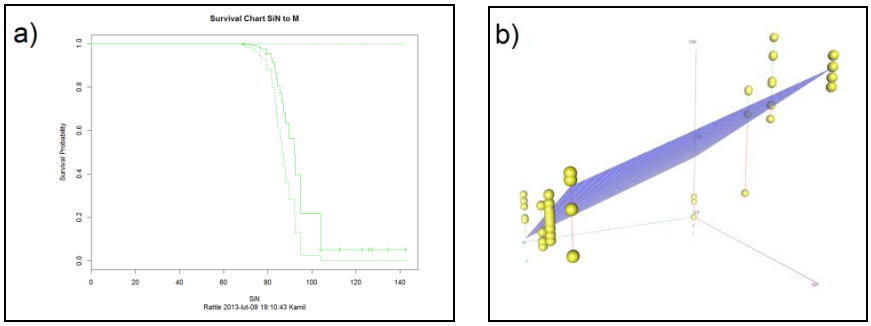


Fig. 1. Graphs: a) survival analysis and b) 3-D.

Efficiency Analysis of Companies Using DEA Model with *R*

Youngchul Shin[1,*]

1. Dept. of Digital Economics, Daejin University in Korea
*Contact author: ycshin@daejin.ac.kr

Keywords: Efficiency, DEA(data envelopment analysis), CCR, BCC, Malmquist

DEA (data envelopment analysis) is a linear programming-based technique proposed by Charnes et al. (1978), which can be used to determine the efficiency of a group of decision-making units (DMUs) relative to an envelope (efficient frontier) by optimally weighting inputs and outputs. Additionally, DEA provides a single indicator of efficiency irrespective of the number of inputs and outputs.

The original model developed by Charnes, Cooper and Rhodes(CCR model) was applicable when characterized by constant returns to scale(CRS). Imperfect competition may cause a DMU not to operate at optimal scale. Banker, Charnes and Cooper(BCC model, 1984) extended the CCR model to account for technologies that show variable returns to scale(VRS). The technical efficiency score (in both CRS and VRS models) equal one implies full efficiency. On the other hand, if the score is less than one it indicated technical inefficiency.

The CCR and BCC efficiency of companies are applied to evaluating the relative efficiency of these companies , if *R* function `dea` is used. The Malmquist productivity indexes of companies are measured and those are decomposed into technical efficiency change and technological change, if *R* function `Dt2t2`, `Dtt2` and `Dt2t` are applied. This analytic method could be used to analyze the efficiencies of companies in the other country.

References

Banker, R.D., Charnes, A., Cooper, W.W.(1984), Some models for estimating of technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9): 1078-92.

Charnes, A., Cooper, W.W. and Rhodes, E.(1978), Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6): 429-444.

Fare R, Grosskopf S, Lindgren B, Ross P.(1994), *Productivity developments in Swendish hospital: a Malmquist output index approach*, Boston:Kluwer, 1994, 253-272.

kisvalue website: www.kisvalue.com

Introducing statistics and probability concepts with R in engineering grades

Pilar Sanmartin*

Department of Applied Mathematics and Statistics. Technical University of Cartagena.
*Contact author: pilar.sanmartin@upct.es

Keywords: Teaching, Simulation, Stochastic Processes

Telecommunications engineering degree offered by the Technical University of Cartagena includes in its program a statistics and probablity course. This subject provides an introduction to probability theory, stochastic processes, queuing systems and basic methods in statistical inference. Laboratory sessions are developed using *R*. This work shows how the use of *R* allows to introduce in a very intuitive way a wide range of concepts. The flexibility of *R* in order to write functions is exploited and complemented by existing functions in *R* packages as **gRbase** and **stats** among others. Some practical examples used in the course are presented.

References

Crawley, M.J. (2007) The R book. Wiley and sons.

Højsgaard, S. Edwards, D. and Lauritzen S. (2012). Graphical models with R. Springer (UserR! series)

Yates, R. and Goodman, D. (2008). Probability and Stochastic Processes. Wiley and sons.

Ross, S. (2008) Simulation. Academic Press.

Biomarker Discovery using Metabolite Profiling Data: Discussion of different Statistical Approaches

Sandra González Maldonado[1,*], Erik Peter[1], Ann-Kristin Petersen[1], Philipp Schatz[2], Oliver Schmitz[1], Henning Witt [1], Jan Wiemer [1]

1. metanomics GmbH, A BASF Plant Science company (Tegeler Weg 33, 10589, Berlin, Germany)
2. Metanomics Health GmbH (Tegeler Weg 33, 10589, Berlin, Germany)
*Contact author: Sandra González Maldonado (sandra.gonzalez-maldonado@metanomics.de)

Keywords: Classification, Biomarker, Metabolite Profiling, Metabolomics, Feature Selection, Omics, Performance Estimation

Metabolite Profiling, the analysis of biochemical pathways that involve small molecules, is a rapidly emerging discipline that helps answering a wide range of biological questions. One of its greatest potentials is the development of criteria that allow us to diagnose a disease or to classify patients according to their risk for a clinical outcome of interest. Such criteria, known as Biomarkers, are commonly used in the medical field.

Metabolite profiling data and omics data in general, pose several statistical challenges. These include: (a) the necessity to adequately account for confounders such as age, BMI, gender, medication and clinical center, by using an appropriate study design in the presence of constraints imposed by clinical praxis, (b) the high ratio of the number of variables to sample size, (c) the required data preprocessing and normalization, (d) the need to generate models as simple, transparent and robust as possible by feature selection and suitable classification methods, and (f) conduct unbiased performance estimation.

The focus of this poster is to give an overview of our statistical approaches and, in particular, to discuss different classification models such as Random Forest or Penalized Logistic Regression, (**randomForest**, **glmnet, penalized)**, feature selection strategies such as recursive feature elimination and forward selection, as well as overall performance estimation. We illustrate our approach by presenting results from one of our recent biomarker studies.

edeR: Email Data Extraction using R

Jaynal Abedin[1,*] Kishor Kumar Das[1], M. A. Yushuf Sharker[1]

1. icddr,b, GPO Box 128, Dhaka 1000, Bangladesh
*Contact author: joystatru@gmail.com

Keywords: Email data extraction, IMAP

The examination of email data for forensics purposes, for understanding underlying organizational network structures and personal email networks are common real-world applications of network analysis. The extraction of email data is required for this type of analysis, and to date *R* cannot extract such data using either Internet Message Access Protocol (IMAP), Simple Mail Transfer Protocol (SMTP), or Post Office Protocol 3 (POP3). The objective of the **edeR** package is to extract email data using IMAP, which will allow users to securely submit their username and password and then extract email data including "to", "from", and "cc" addresses, "date/time", "subject" and email message body. This package will also allow users to search and extract email using specific search tags from an email subject and/or message body and/or can extract messages within a certain date range.

Reproducible and Standardized Statistical Analyses using R

Klaus Marquart[1*], Daniel Chamrad[1], Carmen Theek[1]

1. Protagen AG, Otto-Hahn-Straße 15, 44227 Dortmund, Germany
*Contact author: klaus.marquart@protagen.com

Keywords: statistical analyses, reporting, reproducibility

Currently, SAS is the unchallenged standard for statistical analyses in a biomedical environment. Unfortunately, especially for small companies this is correlated with immense costs. The R environment for Statistical Computing is providing a wide range of statistical methods and graphics, is highly extensible and is free software.

We have developed a setting in which *R* is used in a qualified fashion, supporting some regulatory requirements for validated systems, especially focusing on reproducibility and standardization. As IDE for code development RStudio is chosen and all code is held under version control via Git. Statistical programming in *R* is carried out in a standardized way providing standard statistical analysis reports (SAR) in PDF-format.

Work within projects is organized using a project template with a pre-defined structure of files and folders. Projects are divided in three parts, namely “Data preprocessing”, “Statistical analysis” and “Report generation”.

Statistical analyses use *R* code stored within an in-house developed code library consisting of packages and source files written following software development standards. The statistical analysis plan (SAP) provides details about raw data and the statistical analyses to be performed. The structure of a SAR is based on a tabular overview of tables, figures and listings which is defined in the SAP providing information about hierarchy, numbering, sections and headings of planned analyses. This overview is also available in CSV-format and is included within the report generation process. As result a statistical report in PDF-format is produced as well as a CSV-file for each table or listing and a graphic in PDF-format for each figure in the report. Additionally, results are stored in *R* objects. All reporting functionality is provided by an in-house developed report package, internally using LaTeX and Sweave.

hwriterPlus: Extending the hwriter Package

David Scott[1,⋆]

1. University of Auckland
⋆Contact author: d.scott@auckland.ac.nz

Keywords: Reproducible research, HTML, Microsoft Word

The *R* package **hwriter** provides a convenient way of producing hypertext markup language documents which incorporate statistical analyses using *R*. This package extends the capability of **hwriter** to allow the incorporation of scalable vector graphics, output from *R* or complete *R* sessions and the display of mathematical expressions using LaTeX format. The resulting documents may be successfully viewed in recent versions of common browsers. Additionally such documents may also be opened in recent versions of Microsoft Word.

Application of the nearest neighbour indices in spatstat *R* package for Persian oak (*Quercus brantii* var. *persica*) ecological studies in Zagros woodlands, Iran

Yousef Erfanifard[1,*], Laya Zare[2]

1. Assistant Prof., College of Agriculture, Shiraz University, Shiraz, Iran
Postal Code: 7144165186, Tel/Fax: +98-711-2287159
2. M.Sc. candidate, College of Agriculture, Shiraz University, Shiraz, Iran
*Contact author: erfanifard@shirazu.ac.ir

Keywords: Nearest neighbour indices, Persian oak, Spatial ecology, **Spatstat**, Zagros

Nearest neighbour indices applied in spatial ecology of forest stands describe correlations among trees relative to their distances. They are sensitive to the nature of the pattern of trees, which is extremely variable depending on environmental factors. These indices are well established in the *R* package **spatstat**. We aimed to investigate how the nearest neighbour indices in this package explain the observed pattern of Persian oak coppice trees in Zagros woodlands, Iran. We first took a census in a 9 ha homogeneous plot in these woodlands purely covered with Persian oak trees and tagged all of them to obtained their point map. Spatial pattern analysis of these trees was performed by nearest neighbour indices of *G(r)*, *F(r)*, *J(r)*, neighbourhood density function *(L_x)* and Clark & Evans (*CE)* to recognize the significant deviation of their pattern from CSR. Results showed that the overall spatial pattern of Persian oak coppice trees departed significantly from CSR. There was spatial segregation in their dispersion as indicated by the nearest neighbour indices. The *G(r)*, *F(r)* and *J(r)* indices were clearly above the confidence envelopes of CSR computed by 99 Monte Carlo simulations. Similar results were obtained by L_x and *CE*. Although each one of them explained the spatial dispersion of Persian oak trees differently, *J(r)* could show the regularity of the trees in different distances very well and the map produced by L_x could show the location of dense and sparse areas covered with Persian oak trees. It also was concluded that Persian oak coppice trees were located regularly and they were not related ecologically. These trees were independent and did not affect the establishment of each other that might be because of their coppice structure. This case study shows the efficiency of the *R* package **spatstat** to be implemented in precise and accurate computation of the nearest neighbour indices contributed to ecological studies of forests and woodlands.

References

Baddeley A. and R. Turner (2012). Package 'spatstat'. http://www.spatstat.org

Baddeley, A. and Turner, R. (2005a) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software 12*, 1-42.

Getis A. and J. Franklin (1987). Second-order neighbourhood analysis of mapped point patterns. *Ecology 68*, 473-477.

Illian J., Penttinen A., Stoyan H. and D. Stoyan (2008). Statistical analysis and modelling of spatial point patterns. John Wiley & Sons, UK.

Pommerening A. and D. Stoyan (2008). Reconstructing spatial tree point patterns from nearest neighbor summary statistics measured in small subwindows. *Canadian Journal of Forest Research 38*, 1110-1122.

Stoyan D. (2006). On estimators of the nearest neighbor distance distribution function for stationary point process. *Metrika 64*, 139-150.

Stoyan D. and H. Stoyan (1994). Fractals, random shapes and point fields. John Wiley & Sons, UK.

Point Process Spatio-Temporal Product Density Estimation in R

F. J. Rodríguez-Cortés[1], **J. Mateu**[1], **M. Ghorbani**[2] and **J. A. González**[1]
[1] Department of Mathematics, Universitat Jaume I, Campus Riu Sec, Castellón, Spain
[2] Department of Mathematics Sciences, Aalborg University, Aalborg, Denmark

Abstract

There is an extensive literature on the analysis of point process data in time and in space. However, methods for the analysis of spatio-temporal point processes are less well established. Many spatial processes of scientific interest also have a temporal component that may need to be considered when modelling the underlying phenomenon. Spatio-temporal point processes, rather than purely spatial point processes, must then be considered as potential models. A natural starting point for analysis of spatio-temporal point process data is to investigate the nature of any stochastic interactions among the points of the process. Second-order methods provide indeed a natural starting point for such analysis. In this work kernel estimates of the second-order product density of a spatio-temporal point process with and without considering first- and second-order spatio-temporal separability are given. Further, the expectation and variance of these estimators are obtained, and an approximation of the estimation variances of these estimators for planar Poisson processes is obtained. We finally present a simulation study and an application in the environmental field making use of the library `stpp` and connected Fortran subroutines to R.

Keywords and Phrases: Point processes, spatio-temporal separability, second-order product density, second-order intensity-reweighted stationarity.

Spatio-Temporal ANOVA for replicated point patterns using R

[1], **J A. González**[1] **J. Mateu**[1] and **F. J. Rodríguez-Cortés**[1]
[1] Department of Mathematics, Universitat Jaume I, Campus Riu Sec, Castellón, Spain

Abstract

Spatio-temporal point processes stand as a powerful tool to treat mathematically the idea of randomly distributed points with coordinates corresponding to spatial and temporal components. It has not been paid attention to the analysis of patterns for spatio-temporal replicated data.

There are several approaches when treating spatial replicated data that served as a basis for further statistical analysis. But the analysis of spatio-temporal patterns is not yet fully developed. Large sets of questions could be answered by analysing the underlying structures of the patterns sets. We use R packages as abind and stpp in order to calculate one functional descriptor of pattern for each subject to investigate departures from completely spatio-temporal random patterns. The distributions of our main functional pattern descriptor and of our proposed statistical test are unknown. For nonparametric inference, we use a bootstrap procedure in which we resample our data to estimate the null distribution of our statistical test. A simulation study build in R, provides evidence of the validity and power of our procedure for bootstrap hypothesis testing in our context. We conclude with an application of our method to an ecological dataset of most abundant tree species in the Barro Colorado Island.

Keywords and Phrases: Point processes, spatio-temporal patterns , second-order spatio-temporal properties, ANOVA, replicated patterns.

Estimation of parameters using several regression tools in sewage sludge by NIRS.

L. Galvez-Sola[1] , J. Morales[2], AM. Mayoral[2], C. Paredes[1] , MA. Bustamante[1], FC. Marhuenda-Egea[1], X. Barber[2], R. Moral[1]

1. Department of Agrochemistry and Environment , Miguel Hernandez University. Depart
2. Center of Operations Research, Miguel Hernandez University
*Contact author: xbarber@umh.es

Keywords: NIRS; Biosolids; Partial least square regression (PLSR); Penalized signal regression (PSR); Heavy metals

Sewage sludge application to agricultural soils is a common practice in several countries in the European Union. Nevertheless, the application dose constitutes an essential aspect that must be taken into account in order to minimize environmental impacts. We use, near infrared reflectance spectroscopy (NIRS) to estimate in sewage sludge samples several parameters related to agronomic and environmental issues. Two regression models were fitted: the common partial least square regression (PLSR) and the penalized signal regression (PSR), by **pls** and **mgcv**. Using PLSR, NIRS became a feasible tool to estimate several parameters with good goodness of fit, such as total organic matter, total organic carbon, total nitrogen, water-soluble carbon, extractable organic carbon, fulvic acid-like carbon, electrical conductivity, Mg, Fe and Cr, among other parameters, in sewage sludge samples. For parameters such as C/N ratio, humic acid-like carbon, humification index, the percentage of humic acid-like carbon, the polymerization ratio, P, K, Cu, Pb, Zn, Ni and Hg, the performance of NIRS calibrations developed with PLSR was not sufficiently good. Nevertheless, the use of PSR provided successful calibrations for all parameters.

References

M. Durban, DJ Lee (2008) Splines con penalizaciones (P-splines): Teoría y aplicaciones. Universidad Pública de Navarra.

P.H.C. Eilers, B.D. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemometr. Intell. Lab. Syst., 66 (2003), pp. 159–174

B.D. Marx, P.H.C. Eilers (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline ApproachTechnometrics, 41 (1999), pp. 1–13.

S. Wood (2006). Generalized additive models: an introduction with R. CRC-Press.

Recipe for the implementation of a population dynamics bayesian model for anchovy: Supercomputing using doMC , rjags and coda *R* packages

Margarita María Rincón[1*] and Javier Ruiz[1]

1. Department of Coastal Ecology and Management, Instituto de Ciencias Marinas de Andalucía, Consejo Superior de Investigaciones Científicas, Avda República Saharaui 2, 11519 Puerto Real, Cádiz, Spain
*Contact author: margarita.rincon@csic.es

Keywords: Anchovy, Population dynamics, Bayesian modelling, rjags, parallel computation

Coupling life-cycles with environmental conditions is necessary when modelling the population dynamics of small pelagic fish. We developed a new bayesian model for anchovy in the Gulf of Cádiz that incorporates natural forcing such as the influence of intense winds and sea surface temperature as well as anthropogenic forcing such as fishing and the regulation of freshwater discharges from the Guadalquivir river. Each forcing (natural and human) has different time scales and the model is designed to suit them by using two time-scales; weekly and monthly for the early stages and the remainder of the anchovy life cycle, respectively.

As a consequence of this double time resolution the computational effort is high and the use of a supercomputer becomes necessary. The Centre of Supercomputing of Galicia (CESGA) provide us with their services but the process of adapting our model, taking advantage of parallel computation wasn't trivial.

The main software used is *JAGS 3.3.0* (Plummer, 2013) but it's not flexible enough to summarize the graphical and data outputs, taking this into account Martyn Plummer developed the **rjags** package (Plummer, 2012) making *R* (R Development Core Team, 2011) the user interface where the bayesian data analysis is done. The computer parallelization of the Markov Chain Monte Carlo (MCMC) chains was provided by the **doMC** package (Revolution Analytics, 2012) and the output analysis and diagnostics for MCMC simulations was provided by the **coda** package (Plummer et al., 2006).

We present a detailed methodology to run a bayesian model in a linux based supercomputing center through *R* giving some results of our population dynamics bayesian model for Anchovy in the Gulf of Cádiz. *R* provides a powerful interface that allows us to easily illustrate the accuracy and goodness of fit of the model using quality graphics and diagnostics.

References

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Revolution Analytics (2012). doMC: Foreach parallel adaptor for the multicore package. R package version 1.2.5. http://CRAN.R-project.org/package=doMC

Martyn Plummer (2012). rjags: Bayesian graphical models using MCMC. R package version 3-9. http://CRAN.R-project.org/package=rjags

Martyn Plummer (2013). JAGS version 3.3. 0 user manual. International Agency for Research on Cancer.

Martyn Plummer, Nicky Best, Kate Cowles and Karen Vines (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News 6*, 7-11.