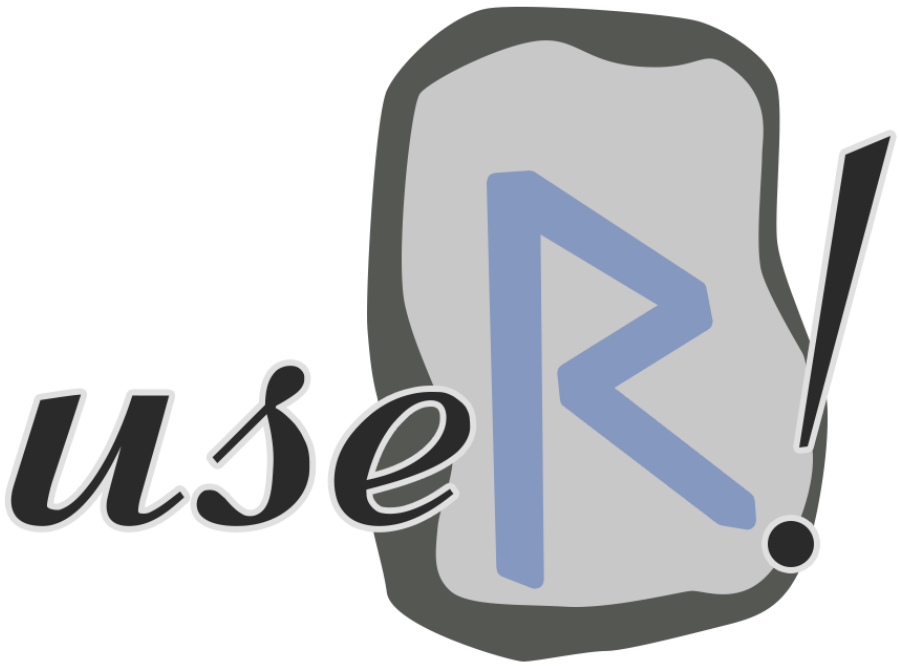
Book of Abstracts




```
> sessionInfo()
[1] "June 30 - July 3, 2015"
[2] "Aalborg, Denmark"
```

June 27, 2015

Conference Sponsors

Diamond Sponsor

Platinum Sponsors

Gold Sponsors

Silver Sponsors

Bronze Sponsors

Media Sponsors

Conference program

Time	Tuesday	Wednesday	Thursday	Friday
08:00	Registration opens	Registration opens	Registration opens	Registration opens
08:30 – 09:00		*Opening session (by Rector peR! M. Johansen, Aalborg University)* Aalborghallen		
09:00 – 10:00	Morning Tutorials *incl. coffee break*	**Romain François** Aalborghallen	**Di Cook** Aalborghallen	**Thomas Lumley** Aalborghallen
10:00 – 10:30		*Coffee break* Sponsored by Quantide	*Coffee break* Sponsored by Alteryx	*Coffee break* (15 min)
10:30 – 12:00		**Session 1** Kaleidoscope 1 Aalborghallen Ecology Gæstesalen Networks Musiksalen Reproducibility Det Lille Teater Interfacing Radiosalen	**Session 4** Kaleidoscope 4 Aalborghallen Medicine Gæstesalen Regression Musiksalen Commercial Offerings Det Lille Teater Interactive graphics Radiosalen	**Sponsor session** (10:15) Aalborghallen DataRobot RStudio Teradata Revolution Analytics alteryx TIBCO H_2O HP
12:00 – 13:00	Sandwiches	*Lunch (standing buffet)* Sponsored by Revolution Analytics	*Lunch (standing buffet)* Sponsored by TIBCO	*Break: 12:00 – 12:30* **Steffen Lauritzen** *(12:30)* Aalborghallen
13:00 – 14:30	Afternoon Tutorials *incl. coffee break*	**Session 2** Kaleidoscope 2 Aalborghallen Case study Gæstesalen Clustering Musiksalen Data Management Det Lille Teater Computational Performance Radiosalen	**Session 5** Kaleidoscope 5 Aalborghallen Teaching 1 Gæstesalen Statistical Methodology 1 Musiksalen Machine Learning 1 Det Lille Teater Visualisation 1 Radiosalen	*13:30:* Closing remarks *13:45:* Grab 'n go lunch *14:00:* Conference ends
14:30 – 15:00		Coffee (with cake)	Coffee (with cake)	
15:00 – 16:00		**Adrian Baddeley** Aalborghallen	**Susan Holmes** Aalborghallen	
16:00 – 17:30		**Session 3** Kaleidoscope 3 Aalborghallen Business Gæstesalen Spatial Musiksalen Database Det Lille Teater *Lightning talks* Radiosalen	**Session 6** Kaleidoscope 6 Aalborghallen Teaching 2 Gæstesalen Statistical Methodology 2 Musiksalen Machine Learning 2 Det Lille Teater Visualisation 2 Radiosalen	
Evening	*Welcome reception* Sponsored by Teradata *(19:00 – 22:00)*	*Poster session, drinks and buffet* Sponsored by DataRobot *(18:00 – 21:00)*	*Conference dinner* Sponsored by RStudio *(18:00 – 23:00)*	

Contents

Part I

Invited speakers

Invited speakers

My R adventures

Romain François

R Enthusiast

http://www.r-enthusiasts.com

Abstract: I will review my adventure with R, from the early days of getting used and addicted to the language and its community, my first attempts at participating in the community through mailing lists and the graphics gallery, valuable collaborations with other members of the community.

For some time now, my interests revolve around expressiveness and performance. The way R lazily evaluates expressions and allows for non standard evaluation has opened a few doors over the years, e.g. the rJava higher level syntax with J, to which I played a humble part, or lately the development of the vocabulary of dplyr. About performance, I've been spending significant time for a few years about the ease of which we can connect R with C++, the development of modern Rcpp has fed my need for both performance and expressiveness. The success of the Rcpp family of packages is a great testimony to efforts that were put in both these directions.

Each of these projects I'm involved with are not personal developments, but rather results of collaboration with relevant members of the community. I have learned a lot from these projects and hopefully shared some of my gained knowledge and expertise along the way.

Finally, I will tackle the exciting new avenues to explore. RcppParallel with its use of the Intel Thread Building Blocks library provides interesting means to approach parallelization with a compelling syntax and set of very approachable design patterns. I will give hints on how this is going to influence further developments in projects I'm mostly involved with, starting from dplyr, Rcpp11, Rcpp14.

How R has changed spatial statistics

Adrian Baddeley

Centre for Exploration Targeting, University of Western Australia

http://www.cet.uwa.edu.au/

Abstract: The growth of R has triggered a revolution in the science of analysing spatial data – especially its 'problem child', the analysis of spatial point patterns. I will sketch some of the revolutionary ideas (past, current and future) and demonstrate them using the contributed package 'spatstat', which played a prominent role in the transformation.

A Survey of Two Decades of Efforts to Build Interactive Graphics Capacity in R

Di Cook

Department of Econometrics and Business Statistics, Monash University
http://www.buseco.monash.edu.au/ebs/

Abstract: What's the difference between these two software descriptions?

"Lisp-Stat is an extensible statistical computing environment for data analysis, statistical instruction and research, with an emphasis on providing a framework for exploring the use of dynamic graphical methods". Luke Tierney, 1998

"R is a programming language and software environment for statistical computing and graphics". Wikipedia, Mar 2015

R has been eighty-seven steps forward for data analysis but twelve steps backwards for interactive graphics, from where XLispStat (and perhaps Data Desk) had put the field in the 1990s. In the intervening twenty years we have seen numerous contributions towards building the same capacity into R, with no absolute success as of yet. In this talk we will describe some of the key package developments, e.g. gwidgets, tcltk2, RGtk2, iplots, rggobi, rgl, SVGAnnotation, and the current exciting ventures, e.g. cranvas, animint, gridSVG, shiny, ggvis. Differences in how the attempts achieve interaction, that illustrate the strengths and weaknesses of the different approaches, will be discussed, with some emphasis on conceptual models for interactive graphics that support data analysis.

Multitype data integration : challenges from the Human Microbiome

Susan Holmes

School of Medicine, Stanford University

med.stanford.edu

Abstract: Using the flexibility of multicomponent objects, we have developed phyloseq: a Bioconductor package for joint analyses of phylogenetic trees, species contingency tables, community graphs and clinical data. I will show examples of reproducible research done on bacterial communities relevant for predicting preterm birth and resilience after perturbations using phyloseq together with standard ecological packages such as vegan and ade4. Finally, more recently we have developed Shiny-Phyloseq which runs as a browser application enabling biologists unfamiliar with R to analyze their microbiome data.

This is joint work with Joey McMurdie and Ben Callahan.

How flexible computing expands what an individual can do

Thomas Lumley

Department of Statistics, University of Auckland

https://www.stat.auckland.ac.nz

Abstract: Design-based inference in survey statistics involves reimplementing everything to allow for unequal sampling probabilities and correlation in sampling, for moderately large data sets. Since there is only one way sampling can be iid, but many ways it can be unequal and correlated, there is a risk of reimplementing everything many times. I will talk about how the survey software in R has taken advantage of advanced concepts such as column-store databases and sparse matrices, and simple concepts such as objects. The result is arguably more comprehensive than any other design-based inference system available, but from at least an order of magnitude less effort.

Linear estimating equations for Gaussian graphical models with symmetry

Steffen L. Lauritzen

Department of Mathematical Sciences, University of Copenhagen
http://www.math.ku.dk

Abstract: In models of high complexity, the computational burden involved in calculating the maximum likelihood estimator can be forbidding. Proper scoring rules such as the logarithmic score, the Brier score, and others, induce natural unbiased estimating equations that generally lead to consistent estimation of unknown parameters. The logarithmic score corresponds to maximum likelihood estimation whereas a score function introduced by Hyvärinen (2005) leads to linear estimation equations for exponential families, including Gaussian graphical models with symmetry.

We shall briefly review the facts about proper scoring rules and their associated divergences, entropy measures, and estimating equations, and show how Hyvärinen's rule leads to simple estimating equations for Gaussian graphical models. Finally, it shall be discussed how these estimates can be used for fast model selection in Gaussian graphical models.

Reference P. G. M. Forbes and S. Lauritzen. Linear Estimating Equations for Exponential Families with Application to Gaussian Linear Concentration Models. (2014). Linear Algebra and its Applications, 473: 261–283.
doi: 10.1016/j.laa.2014.08.015 see also arXiv:1311:0662

Part II

Oral Presentation

Kaleidoscope 1

Chair: Heather Turner

flowcatchR: A user-friendly workflow solution for the analysis of time-lapse cell flow imaging data

Federico Marini

Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), Division Biostatistics and Bioinformatics, University Medical Center of the Johannes Gutenberg University Mainz, Germany

http://www.imbei.uni-mainz.de

Abstract: Automated bioimage analysis is required for reproducible and efficient extraction of information out of time-lapse microscopy data when investigating the in vivo dynamics of complex biological processes.

We developed a comprehensive workflow solution, based on our R/Bioconductor package flowcatchR, the first capable of effectively handling cell tracking in R. Our solution specifically addresses blood cell flow data, where cells show dynamic behaviors (flowing, rolling). Analysis of the corresponding fast movements is further complicated by cells entering and leaving the field of view, and transitions in and out of focus.

Subject matter knowledge is incorporated for making analysis feasible. Specifically, we developed a penalty function for a tracking algorithm to account for the directionality of the flowing cells.

Our workflow solution, based on an R package implementing the algorithms, a Shiny App, and Jupyter notebooks, also serves as a proposal for bridging the gap between sophisticated analysis tools and end-user requirements in bioimaging applications.

Additionally, we use custom-made Docker containers for efficient deployment, for providing subject matter researchers with fully operative environments, where all necessary libraries and dependencies are pre-installed and ready for use. This approach can guarantee high levels of reproducibility while being accessible to a broad range of scientists.

Keywords: Bioimaging, Cell tracking, Bioinformatics, Workflow, Automated analysis

Image processing and alignment with RNiftyReg and mmand

Jonathan Clayden

Institute of Child Health, University College London, UK

http://www.ucl.ac.uk/ich/

Abstract: Images are produced in a wide spectrum of science and engineering disciplines, and in some fields they may have three or more dimensions. A range of postprocessing techniques are commonly used to identify or emphasise features of interest. It is also often necessary to align pairs of images, a process otherwise known as registration. Here we will present two packages, written in R and C/C++, which are designed to be used in this domain. RNiftyReg performs linear and nonlinear registration between 2D or 3D images, and allows estimated transformations to be applied to other images and points. The mmand package (for "mathematical morphology in any number of dimensions") offers facilities for image erosion and dilation, smoothing and filtering. It can also be used for other, related kernel-based operations, and for array resampling or noninteger indexing.

Keywords: image processing, image registration, mathematical morphology

rags2ridges: Ridge estimation and graphical modeling for high-dimensional precision matrices

Carel F. W. Peeters

Dept. of Epidemiology & Biostatistics, VU University medical center Amsterdam, the Netherlands

http://www.vumc.nl/afdelingen/EB/

Abstract: Estimation of the multivariate normal precision matrix is central to many statistical procedures. We study ridge estimation of the precision matrix in the high-dimensional setting where the number of variables is large relative to the sample size. We note that several classes of estimator that are caught under the umbrella term 'ridge-type precision estimation' cannot be explained as resulting from penalization with a common L2-penalty. Subsequently, starting from a proper L2-penalty, analytic expressions are derived for two alternative ridge estimators (encapsulating target and non-target shrinkage) of the precision matrix.

The rags2ridges package implements the alternative ridge estimators along with supporting functions to employ these estimators in a graphical modeling setting. These supporting functions enable, a.o., the determination of the optimal value of the penalty parameter, the determination of the support of a shrunken precision estimate, as well as various visualization options.

We will demonstrate some of the properties of the alternative precision estimators and show how the R implementation can be used in the graphical modeling of oncogenomics data. In addition, we will present some of the modules in rags2ridges that are currently under development. These modules extend the basic module to deal with meta-analytic graphical modeling and directed networks.

Keywords: graphical modeling, high-dimensional precision matrix estimation, L2-penalization, multivariate normal, network inference from oncogenomics data

dgRaph: Discrete factor graphs in R

Henrik Tobias Madsen

BiRC and MOMA, Aarhus University, Denmark

http://birc.au.dk/

Abstract: Factor graphs provide a flexible and general framework for specifying probability distributions. Well-known models such as Hidden Markov Models and Mixture models can be cast into the framework yielding a graph structure that easily reveals the dependency structure.

dgRaph is an R-package for inference and significance evaluation in discrete factor graphs.

Factor graphs can be specified, plotted and manipulated using familiar R data structures and functions. We implement some standard algorithms; sum-product to infer likelihood and posterior distribution of hidden variables, max-sum for inferring most probable state of hidden variables, the EM-algorithm for inferring parameters.

Furthermore dgRaph contains two novel methods for comparison of a null and a foreground model, with identical graph structure but different sets of potentials. We evaluate the probability of an observation with larger log-odds score under the null-model. The first method is importance sampling where the importance sampling distribution is automatically chosen. The second is employing saddlepoint approximation.

The R-package binds a C++-library using Rcpp. Thus carrying out the time-consuming recursive calculations in low-level, compiled code, still having the interactivity and flexibility of R.

dgRaph will be released soon to CRAN until then it can be installed from [github].

Keywords: Graphical models, Statistical modelling, Significance evaluation

Ecology

Chair: Hadley Wickham

Optimized R functions for analysis of ecological community data using the R virtual laboratory (Rvlab)

Costas Varsos and Theodore Patkos

Institute of Computer Science, Foundation of Research and Technology Hellas, Heraklion, Crete, Greece

http://www.ics.forth.gr/index.html

Abstract: Parallel data manipulation is no news to the R community; yet, it is not uncommon even for experienced developers not to have a clear recipe as to which package to select among the repertoire of choices for handling Big Data in R. Our targeted users, the ecologists/microbiologists/average R users, often face difficulties in exploiting the full capacity of their computational resources to execute their R scripts. This report has a twofold goal: (i) describe a complete methodology for the analysis of large datasets, combining diverse R packages, (ii) present its application on a virtual R laboratory (RvLab) that makes execution of complex functions and visualization of results, easy and readily available to the end-user.

Our methodology processes data at different levels of abstraction. The pbdMPI package is used to implement Single Program Multiple Data (SPMD) parallelization on primitive mathematical operations (e.g., outer product), as well as on their interplay within popular high-level functions of the Vegan package. The dplyr and RPostgreSQL packages are further integrated, offering secondary storage solutions whenever memory demands exceed RAM resources.

RvLab is running on a PC cluster and offers an intuitive UI enabling the analysis of ecological and microbial communities based on optimized Vegan functions.

Keywords: Ecological Community Data, Parallelization, Vegan, R Virtual Lab, Big Data

Building ecological models bit-by-bit

David L Miller

Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, Scotland

http://creem2.st-andrews.ac.uk/

Abstract: Many models that we build rely on combining several components, often representing different processes at work in the system. For example in ecology we may have a component for detectability of the species by observers, one for whether the animals are available, one describing their spatial distribution.

Defining such a complex model in R can result in writing a very long call to a particular function, which may or may not work at any point. Since these processes are often (conditionally) independent, we can separate these components in the likelihood and code. Using the + operator we can then combine them in a simple way that's easy for those doing the modelling to understand. Since each step produces an object, we can perform model validation, diagnostics and checking during construction.

Through an example applied to distance sampling, I'll show an example of this kind of approach to model building and propose some other situations where this kind of strategy may also be useful. I hope this will provoke some debate on the user friendliness of modelling packages and their interfaces.

Keywords: ecology, modelling, user interface, distance sampling

Simulating ecological microcosms with systems of differential equations: tools for the scientific, technical and communication challenges.

Andrew Dolman

Dept. of Freshwater Conservation, Brandenburg University of Technology Cottbus - Senftenberg, Bad Saarow, Germany

http://www.tu-cottbus.de/fakultaet4/en/freshwater-conservation/

Abstract: To better understand complex natural systems, ecologists combine experiments on microcosms with simulations using systems of ordinary differential equations (ODEs). Good communication is required between theorists, modellers and experimentalists; and computational performance becomes an issue for parameter estimation by maximum likelihood or MCMC. Here we present a workflow in R that addresses these challenges.

The rodeo package allows ODE models to be written as plain text tables, binding mathematical model components with documentation, but separating them from platform specific code – improving communication between modellers and theorists.

rodeo then generates a stoichiometric matrix form of the model in either R, or FORTRAN95 code. Matrix notation eliminates many redundant terms from the ODE, and FORTRAN provides greater computational performance.

The ODE is integrated numerically using solvers provided by the deSolve package, which allows experimental manipulation of the simulated microcosms via events and forcings.

FME provides functions fitting the model to microcosm data via maximum likelihood and MCMC.

Finally, Shiny apps reveal the behaviour and capabilities of the models and greatly enhance communication between modellers and experimentalists.

We illustrate this workflow by modelling a series of nutrient addition bioassays, microcosm experiments used to determine the nutrient limiting phytoplankton growth in lakes.

Keywords: ecology, dynamic modelling, microcosms, ordinary differential equations (ODE)

A Graphical User Interface for R in an Integrated Development Environment for Ecological Modeling, Scientific Image Analysis and Statistical Analysis

Marcel Austenfeld

eLK.Medien, University of Kiel

https://elearning.uni-kiel.de

Abstract: For the exploration and analyzing of natural systems many different tools have to be used to derive hypotheses about causal relationships of complex patterns and processes.

Model tools can help to analyze uncertain quantities, Image analysis tools to measure real patterns and statistical analysis tools for the structuring, interpretation and presentation of simulated and measured data.

The open source application Bio7 was developed to offer these tools beside a selection of popular programming languages and several Graphical User Interfaces in an Integrated Development Environment based on the Eclipse Rich Client Platform.

This talk provides an overview of the Bio7 R interface containing an advanced R script editor with a debugging interface, a spreadsheet component, some special plotting features and other useful tools for development.

In addition some unique and easy to use methods are presented to transfer complex image data from an embedded image interface based on the scientific image analysis software ImageJ.

Finally this presentation should also demonstrate the grown usefulness of the Bio7 tools for related disciplines dealing with simulation, image or data analysis.

Keywords: Ecological Modeling, Eclipse Rich Client Platform, Graphical User Interface, ImageJ, Scientific Image Analysis

Networks

Chair: Claus Dethlefsen

fbRads: Analyzing and managing Facebook ads from R

Gergely Daroczi

CARD.com, United States of America

http://card.com

Abstract: Facebook made its ads marketing and ads API generally available in March of 2015, which empowers developers to analyze, report on and manage marketing actions in a programmatic and automated way. As the API is now generally available to the public, CARD.com decided to release its related R codebase in the means of an R package, which facilitates rapid development around data-driven ad management.

The core features of the package includes support for multiple FB accounts, batch queries and automated paging for larger amount of data, query threshold and basic error handling. Besides a number of wrapper functions around the FB marketing API, the package is intended to be modular enough to support other API endpoints as well.

This talk will be the first publicly available introduction on the package features and how to use it, which is to be soon accompanied by a more detailed vignette based on the slides of the talk and feedback collected at the conference.

Keywords: API, facebook, marketing, adtech, Web Technologies

Web scraping with R - A fast track overview.

Peter Meißner

Department of Politics and Administration, University Konstanz, Germany
http://www.polver.uni-konstanz.de/en/department-home/

Abstract: The presentation is thought to give an overview on how to best approach the task of collecting data from the web with the R programming environment and its packages. While web techniques like HTML, XML, JSON, HTTP, JavaScript, web APIs, ... are mainly easy to understand as such, their abundance as well as a literature either written for computer scientists or web designer but seldom for those 'only' interested in getting the data, pose serious hurdles on scientists in need to collect data but lacking more technical backgrounds. The presentation aims at providing a framework to understand web techniques and the process of conducting data gathering within the web. Furthermore relevant 'best practice' R packages and templates will be introduced along the way.

Keywords: web scraping, web data collection, web data extraction

multiplex: Analysis of Multiple Social Networks with Algebra

Antonio Rivero Ostoic

BADM, Aarhus University, Denmark

http://badm.au.dk/

Abstract: multiplex - Analysis of Multiple Social Networks with Algebra is a package for the study of social systems made of different types of relationships and actors having multiple affiliations. With multiplex is possible to create and manipulate multivariate network data with different formats, and there are effective ways available to treat multiple networks with routines that combine algebraic systems like the partially ordered semigroup and semiring structures for signed graphs together with the relational bundles occurring in different types of multivariate network data sets. The newest version of the package supports Galois connections between families of subsets such as the actors in a social network and their corresponding events. In conjunction with Rgraphviz, there is a routine to visualize bipartite graphs and lattice diagrams of partially ordered sets.

Keywords: relational algebra, multiple networks, algebraic models

What's new in igraph and networks

Gabor Csardi

Department of Statistics, Harvard University, Cambridge, MA, USA
http://statistics.fas.harvard.edu

Abstract: igraph is the premier R package for the analysis of network data and it went through major restructuring recently and has changed a lot since last time it was featured at useR! in 2008.

This talk introduces the new/updated features of igraph:

- Simplified ways of graph manipulation.
- New methods community detection.
- New layouts for graph visualization.
- New statistical methods: graphlets, embeddings, graph matching, cohesive blocks, etc.
- How to use igraph graphs with new visualization tools: DiagrammeR, D3, etc.

Keywords: Network data, igraph, graphs

Reproducibility

Chair: Martin Maechler

rOpenSci: A suite of reproducible research tools in R

Karthik Ram

University of California, Berkeley & The rOpenSci project
http://ropensci.org

Abstract: The rOpenSci project (http:/ropensci.org) began as a grassroots effort in 2011, and has been operating as a fully funded research project since 2013. This community driven project builds upon R's popularity as a scientific research tool and fills in additional gaps in the computational pipeline by providing tools for data acquisition (from APIs and other source), data manipulation (including an emerging suite for spatial data analysis and mapping), data visualization tools, and data publication tools.

In this talk I'll highlight a subset of these tools and how they can facilitate transparent, reproducible and efficient data analysis pipelines.

Keywords: API, R, reproducibleresearch

Enhancing reproducibility and collaboration via management of R package cohorts

Michael Lawrence

Computational Biology, Genentech, South San Francisco, USA

http://www.gene.com/

Abstract: Science depends on collaboration, result reproduction, and the development of supporting software tools. Each of these requires careful management of software versions. We present a unified model for installing, managing, and publishing software contexts in R. It introduces the package manifest as a central data structure for representing version specific, decentralized package cohorts. The manifest points to package sources on arbitrary hosts and in various forms, including tarballs and directories under version control. We provide a high-level interface for creating and switching between side-by-side package libraries derived from manifests. Finally, we extend package installation to support the retrieval of exact package versions as indicated by manifests, and to maintain provenance for installed packages. The provenance information enables the user to publish libraries or sessions as manifests, hence completing the loop between publication and deployment. We have implemented this model across two software packages, switchr and GRANbase, and have released the source code under the Artistic 2.0 license.

Keywords: reproducibility, collaboration, packages

A Review of Meta-Analysis Packages in R

Joshua R. Polanin & Emily A. Hennessy

Peabody Research Institute, Vanderbilt University, Nashville, TN, USA

http://www.vanderbilt.edu/

Abstract: A fast-growing market for meta-analytic software is R. Despite the pervasive availability of meta-analytic R packages, only a small number of publications have examined the abilities of these packages (Chen & Pence, 2013; Neupane, Richer, Bonner, Kibret, & Beyene, 2014), and no review to date has comprehensively cataloged each package. As such, the purpose of this project is to delineate the scope and breadth of meta-analytic R packages in order to: a) inform the meta-analytic field of the availability of these packages, b) explore the packages' power and applicability, c) elucidate future areas of greater package need.

To answer these questions, we conducted a systematic review of meta-analytic R packages using various aggregators (e.g., CRAN website, Revolution Analytics, Crantastic). The search process yielded 136 potential packages of which 54 provided unique contributions (See online appendix). We coded specific capabilities of each package using a pre-defined codebook (See online appendix). Although the coding process is still ongoing, preliminary results indicated a plethora of traditional functionality (e.g., weighted average calculations, moderator analyses), but few packages for the review stage (i.e., screening or coding) or advanced analyses (i.e., multilevel modeling, graphics). The final results will be available prior to the June conference.

Keywords: meta-analysis, review of R packages, needs assessment

Simple reproducibility with the checkpoint package

David Smith

Revolution Analytics (Microsoft)

http://www.revolutionanalytics.com

Abstract: Almost every R script uses packages, but the ever-changing ecosystem of package repositories presents a challenge for reproducible data analysis. It is difficult to use — and especially, to share — R code that depends on packages, when the results may be dependent on specific package versions being used. In this talk I will demonstrate how Revolution R Open and the "checkpoint" package address this problem.

Keywords: reproducibility, packages, checkpoint, revolution r open

Interfacing

Chair: Dirk Eddelbuettel

Some lessons relevant to including external libraries in your R package

Kasper D. Hansen

Department of Biostatistics, Johns Hopkins University, Baltimore, USA

http://www.biostat.jhsph.edu

Abstract: It is tempting - and sometimes necessary - to depend on external libraries of code written by someone else in a non-R language. However, such dependency can be a liability if the external library is not sufficiently well written, especially in terms of portability. I illustrate my experience with this situation with lessons learned from development and maintanence over several years of the Bioconductor packages affxparser and Rgraphviz.

Keywords: Packages, Development, ForeignInterfaces, Respositories

Integrating R with the Go programming language using interprocess communication

Christoph Best

Google, Hamburg, Germany

Abstract: Integrating R into production environments still poses challenges. While there are a number of ways of calling R scripts and libraries from production languages such as C/C++, Java, Python, and others, these are often somewhat fragile. Newer languages like Go which have runtime structures that differ markedly from standard C often do not provide any obvious way to invoke R except by going through the shell.

We have recently explored an approach in which R code executes as an independent, but tightly integrated subprocess of a Go program and communicates over inter-process messages encoded in the protocol buffer format that is widely in use at Google. The R subprocess effectively provides an "execution service" that allows to send data from Go to R, execute R statements, and retrieve the results in a simple, robust, and straightforward way. (similar to the Rserve project, but simpler and more tightly bound to the Go process). We report on our experiences of integrating Go production code with R libraries in this way.

Keywords: cross-language integration, interprocess communication, Go programming language

Naturally Sweet Rcpp with Modern C++ and Boost

Matt P. Dziubinski

Department of Mathematical Sciences, Aalborg University, Denmark
http://personprofil.aau.dk/profil/127800

Abstract: "C++11 feels like a new language: The pieces just fit together better than they used to and I find a higher-level style of programming more natural than before and as efficient as ever." – Bjarne Stroustrup.

Since 2011, Standard C++ has become a simpler, more productive language – it has also expanded its support for numerics and parallelism. This allows us to stay within a portable C++ code while achieving the goals of Rcpp sugar.

This example-driven session will provide a walkthrough showing how modern C++ can be used for a variety of statistical computing applications. We will see how the advances in the core language, the standard library, and the wider libraries ecosystem (particularly Boost) enable simple and efficient code.

Topics include: Data wrangling with C++11 algorithms, lambdas, and Boost (Phoenix, Range, StringAlgorithms, Tokenizer); random number generation with C++11; numerics & statistical analysis with Boost.Accumulators & Boost.Math; easy parallelism with C++11 (async) and OpenMP; timing code with the chrono library; shorter code with auto, decltype & range-based for loop; the most useful algorithms & containers in C++11 and Boost.

We will end with a preview of the upcoming C++14 features which further contribute to a simpler, more readable code.

Keywords: Rcpp, C++11, Boost, Modern C++, Computational Statistics

Linking R to the Spark MLlib Machine Learning Library

Dan Putler

Alteryx, Inc.

http://www.alteryx.com

Abstract: Apache Spark is rapidly emerging as a central technology for addressing large volumes of data. Spark provides a distributed, in-memory processing engine that can be used on top of a Hadoop HDFS data store (also Cassandra and Tachyon). Currently, Spark offers APIs for Scala, Python, and Java, but the next Spark release (1.4) will see the inclusion of an R API to Spark (SparkR). Part of that API involves the creation of a set of bindings to Spark's MLlib library for machine learning. In this presentation, we will discuss three things related to that implantation: (1) the nature of the bindings between R and MLlib, including adapting the emerging ML Pipeline framework to R; (2) the tools and methods these bindings provide to R users; and (3) how concepts and approaches developed for R are influencing Spark in general and MLlib in particular.

Keywords: Spark, High performance computing, Hadoop, R Bindings

Kaleidoscope 2

Chair: Romain François

archivist: Tools for Storing, Restoring and Searching for R Objects

Przemyslaw Biecek

University of Warsaw, Poland

http://www.icm.edu.pl

Abstract: Open science needs not only reproducible research but also accessible final and partial results. The archivist package supports the storing, restoring and searching for an R objects easy. Want to share your object with article reviewers or collaborators? This package should help.

Data exploration and modelling is a process in which a lot of data artifacts are produced. Artifacts like: subsets, data aggregates, plots, statistical models, different versions of data sets and different versions of results. The more projects we work with the more artifacts are produced and the harder it is to manage these artifacts. Archivist helps to store and manage artifacts created in R. Archivist allows you to store selected artifacts as a binary files together with their metadata and relations. Archivist allows to share artifacts with others, either through shared folder or github. Archivist allows to look for already created artifacts by using it's class, name, date of the creation or other properties. Makes it easy to restore such artifacts. Archivist allows to check if new artifact is the exact copy that was produced some time ago. That might be useful either for testing or caching.

Keywords: open science, repository of r objects, r object sharing, archivist, meta data

R User Groups

Joseph B. Rickert

Revolution Analytics / Microsoft

http://www.revolutionanalytics.com/

Abstract: R User Groups continue to thrive around the world. Several new groups were established during the past year and membership in existing groups is growing. In this talk I will examine the dynamics of R user group activity, highlight apparent trends and make some conjectures about the needs that R user groups appear to satisfy and present some ideas about best practices. I will conclude by offering some thoughts on how to start a new user group and describe my experiences helping to grow the Bay Area useR Group.

Keywords: R User Groups, trends, best practices

Computational Precision and Floating-Point Arithmetic: A Teacher's Guide to Answering FAQ 7.31

Richard M. Heiberger

Department of Statistics, Temple University, Philadelphia, PA, USA

http://www.temple.edu

Abstract: Beginners ask questions like "Why does .3 + .6 not equal .9?" Experts always reply "See FAQ 7.31". The answer in the FAQ is correct, yet probably not helpful to the Beginner.

I show several simple arithmetic and algebra statements whose machine-calculated values differ from the values the real number system would provide. I explicitly show the floating-point bit patterns for the numbers, and show why the calculated answer is correct in the floating-point system.

Double precision floating-point numbers use 53 significant bits (about 16 decimal digits). It is difficult for beginners to follow details that depend on behavior in the 53rd bit (16th decimal digit).

I make a pedagogical simplification by using low-precision floating-point numbers (5 significant bits). It is easier to follow unanticipated behavior at 1.5 decimal digits than at 16 decimal digits.

I use the Rmpfr package to work with the 5-bit numbers. While Rmpfr is motivated for increased precision, it works well with reduced precision. Rmpfr automates the apparently simple technique of "rounding at all intermediate steps". Students have trouble detecting "all" intermediate steps and some trouble with correct rounding (round-to-even) at each step.

Keywords: FAQ 7.31, Precision, Floating-Point, Teaching, Rmpfr

Tiny Data, Approximate Bayesian Computation and the Socks of Karl Broman

Rasmus Bååth

Lund university Cognitive Science

http://www.sumsar.net

Abstract: Big data is all the rage, but sometimes you don't have big data. Sometimes you don't even have average size data. Sometimes you only have eleven unique socks. This is the story about how a tweet by esteemed biostatistician Karl Broman resulted in a solution to the old question: How many socks are there in my laundry?

Key to the solution was using R to do approximate Bayesian computation, a conceptually simple technique that allows fitting any generative model to a data set. Approximate Bayesian computation is easily implemented in R using the standard random number generators (rnorm, runif, etc.) and allows for flexible inclusion of prior information, which can be useful when modeling tiny datasets. Using the sock problem as a case study, this talk will introduce modeling of tiny data by approximate Bayesian computation using R.

Keywords: Statistical modelling, Approximate Bayesian Computation, Small data sets

Case study

Chair: Claus Dethlefsen

Using R for small area estimation in the Norwegian National Forest Inventory

Johannes Breidenbach

National Forest Inventory, Norwegian Forest and Landscape Institute
http://skogoglandskap.no

Abstract: The Norwegian National Forest Inventory (NFI) is a national survey program for monitoring the state and development of forests based on more than 22,000 permanent inventory sample plots. One fifth of the plots within forests are visited every year to measure trees. R is an important tool for data analysis in the NFI. For example, all analysis required for the reporting duties arising from the UN Climate Convention and its Kyoto Protocol are based on R. Another important field of application is small area estimation (SAE) for unplanned domains with few sample plots. Auxiliary variables obtained from airborne laser scanning or image matching are used in combination with field sample plots to obtain estimates of biomass or timber volume with acceptable levels of precision. Some methods are implemented in the R package JoSAE. Unit and area-level SAE methods are currently under consideration for future applications and are compared in this study.

Keywords: Small area estimation, Official statistics, forest inventory

Using R for natural gas market balancing in the Czech republic

Ivan Kasanický

Czech Institute of Informatics, Robotics, and Cybernetics, CTU in Prague

http://www.ciirc.cvut.cz/

Abstract: An open gas market, which EU is gradually heading to, requires daily balancing. This is generally a complicated task since most of the end users' consumption is read in longer time intervals (i.e. monthly, annually etc.). Moreover, the reading intervals for the different customers can be different or can be even totally irregular. The network balance input data have therefore various time resolutions from hourly data (network delivery points, large customers) to monthly, annual or even longer period data (smaller customers). We have developed several statistical models of gas consumption using mainly temperature as an explanatory variable. These models can be used for disaggregation of consumption data to the daily values, but also in some cases for forecasting. In the proposed contribution we will demonstrate the whole process of developing the mentioned models, which has been completely done using the R environment. The process consists mainly of large data preprocessing, model selection, and parameter estimation. Since primary users of these models are usually not very familiar with R, we have developed two R packages containing self-explanatory functions for data processing, estimation of model parameters and evaluating the prediction (either disaggregated values or forecast).

Keywords: natural gas consumption , standardized load profiles, r package

Heteroscedastic censored and truncated regression for weather forecasting

Jakob W. Messner

Department of Statistics, University of Innsbruck, Austria
http://www.uibk.ac.at/statistics

Abstract: This contribution presents the crch R-package that provides functions to fit censored or truncated regression models with conditional heteroscedasticity. Maximum likelihood estimation is used to fit Gaussian, logistic, or student-t distributions to left and/or right censored or truncated responses. Different regressors can be used to model the location and the log-scale of these distributions. The functions return S3-objects for which standard methods like print(), summary(), predict(), coef(), vcov(), or logLik() are available.

One application of these models is weather forecasting. Weather forecasts are usually based on numerical weather predictions. However, because of uncertain initial conditions and unresolved atmospheric processes these predictions often exhibit errors. To estimate the forecast uncertainty many weather centers provide ensemble predictions: several numerical predictions that use slightly different initial conditions and model formulations. Because these ensemble predictions can not consider all error sources they are often still uncalibrated and can considerably be improved by statistical models like those provided by the crch package.

With data from the crch package we show that non-negative precipitation can appropriately be modeled with a censored logistic distribution and that the ensemble mean and spread serve as well suited regressors for the location and scale respectively.

Keywords: distributional regression, likelihood regression, heteroscedastic tobit, model output statistics

Multinomial functional regression with application to lameness detection for horses

Helle Sørensen

Department of Mathematical Sciences, University of Copenhagen, Denmark
http://www.math.ku.dk/

Abstract: Our data consists of 85 acceleration signals collected from trotting horses that are either healthy or have an induced lameness on one of the four limbs. Our aim is to develop a method that uses such a signal for detection of lameness and identification of the lame limb. This is a supervised classification problem with five groups and functions (curves) as predictors. We propose to use a multinomial functional regression model. We combine the discrete wavelet transform and LASSO penalization for estimation of the model and use the fitted model to predict the class membership for new curves. We mainly rely on existing R packages for computations (fda for preprocessing of the data signals, wavethresh for wavelet computations, and glmnet for penalized multinomial regression). Joint work with Seyed Nourollah Mousavi.

Keywords: Multinomial functional regression, Supervised classification, Wavelets, LASSO, Lameness detection

Clustering

Chair: Martin Maechler

Unsupervised Clustering and Meta-Analysis using Gaussian Mixture Copula Models

Anders Ellern Bilgrau

Department of Mathematical Sciences, Aalborg University, Denmark
http://math.aau.dk

Abstract: Methods for unsupervised clustering is an important part of the statistical toolbox in numerous scientific disciplines. Tewari, Giering, and Raghunathan (2011) proposed to use so-called semi-parametric Gaussian Mixture Copula Models (GMCM) for general unsupervised clustering when obvious non-gaussian clusters are present. Li, Brown, Huang, and Bickel (2011) independently discussed a special case of these GMCMs as a novel approach to meta-analysis in high-dimensional settings. GMCMs have attractive properties which make them highly flexible and therefore interesting alternatives to well-established methods. However, parameter estimation is hard because of intrinsic identifiability issues and intractable likelihood functions. Both aforementioned papers discuss similar expectation-maximization-like (EM) algorithms as their pseudo maximum likelihood estimation procedure. We have written an improved implementation in R of both classes of GMCMs along with various alternative optimization routines to the EM algorithm. The software is freely available on CRAN through the R-package GMCM. The implementation via RcppArmadillo is fast, general, and optimized for very large numbers of observations.

Keywords: high-dimensional analysis, unsupervised clustering, meta-analysis, copula theory

Hierarchical Cluster Analysis of hyperspectral Raman images: a new point of view leads to 10000fold speedup

Claudia Beleites

Department Spectroscopy and Imaging, Leibniz Institute of Photonic Technology, Jena, Germany

http://www.ipht-jena.de

Abstract: Hierarchical Cluster Analysis (HCA) is an established tool for vibrational spectroscopic data analysis. In general, two modes of clustering are possible: Q-mode HCA clusters spectra (cases) according to their similarity, whereas R-mode analysis clusters variates (spectral bands) accoring to similarity in their distribution profile over the spectra. For hyperspectral images, this corresponds to grouping similar images. Both modes are combined in heatmaps e.g. for microarray data. For vibrational spectra, however, the physical processes generating spectra lead to several spectral bands originating from the same chemical species and thus give a very strong coupling between the results obtained by both modes. Typical hyperspectral imaging data sets nowadays consist of 10^3-10^6; spectra with 10^2-10^3 variates (wavelength bands). Exploiting the physico-chemical relations of the data-generating spectroscopic processes, we reduce the clustering problem to R-mode HCA of a small number of relevant wavelengths, leading to a 10000fold speedup. In addition, R-mode HCA corresponds well with traditional techniques of spectra interpretation, where band assignments yield information about chemical species.

Acknowledgements: CB is funded by BMBF Project RamanCTC (13N12685).

Reference: A Bonifacio, C Beleites & V Sergo: AnalBioanalChem, 2015, 407, 1089-1095. DOI 10.1007/s00216-014-8321-7

Keywords: hierarchical cluster analysis, vibrational spectroscopy, physical chemistry, biospectroscopy, cartilage

Dirichlet process Bayesian clustering with the R package PReMiuM

Silvia Liverani

Department of Mathematics, Brunel University London, UK
http://www.brunel.ac.uk/cedps/mathematics

Abstract: PReMiuM is a recently developed R package for Bayesian clustering using a Dirichlet process mixture models. It is an alternative to regression models, non-parametrically linking a response vector to covariate data through cluster membership (Liverani et al, Journal of Statistical Software, 2015). Posterior inference is carried out by using Markov chain Monte Carlo simulation and to allow for fast computations, all essential methods in the package are based on efficient C++ code.

The model allows binary, categorical, count, survival and continuous response, as well as continuous and discrete covariates. Additionally, predictions may be made for the response, and missing values for the covariates are handled. Several samplers and label switching moves are implemented along with diagnostic tools to assess convergence (Hastie et al, Statistics and Computing, 2014). A number of R functions for post-processing of the output are also provided. In addition to fitting mixtures, it is also possible to determine which covariates actively drive the mixture components.

This talk will include an overview of the features of the package and some of its applications to date.

Keywords: Clustering, Dirichlet process mixture model, Profile regression, Bayesian modelling

Examining the Environmental Characteristics of Tornado Outbreaks in the United States using Spatial Clustering.

Thomas Jagger

Geography, Florida State University, Tallahasse Florida, USA

http://myweb.fsu.edu/jelsner/People.html

Abstract: We determine tornado outbreaks using median clustering with cluster size estimates using pmak() from the flexible procedures for clustering package.

The clusters are determined using each tornadoes touchdown latitude and longitude for each day with at least 16 tornadoes. We use a generalized linear model to examine the relationship between the characteristics of storms within the clusters to the mean environmental variables within the cluster.

For the cluster storm characteristics, we use the average kinetic energy of each tornado in the cluster along with the total number of tornadoes and strong tornadoes. For the environmental conditions, we generate cluster regions using an 80 km buffer around the convex hull of tornado touchdown points within each cluster and calculate the mean convective available potential energy, storm relative helicity and wind shear.

We use R code, and provide reproducible code on R-Pubs using R Markdown and R Presentation from RStudio.

Keywords: Mediod Clustering, Generalized Linear Models, Tornadoes , Reproducible Research, ggplot graphics

Data Management

Chair: Thomas Lumley

Taking testing to another level: testwhat

Filip Schouwenaars

DataCamp

https://www.datacamp.com

Abstract: The architecture of the testthat R package (the de facto standard for writing unit tests for R packages) is very generic and suits itself to extension and adaptation. DataCamp has adapted the testthat package to be used on the R backend of its interactive learning platform. By defining a new type of reporter and adding user-friendly test functions that are designed specifically for testing the correctness of a student's submission, the testwhat package now exists as a wrapper around testthat (https://github.com/Data-Camp/testwhat). The talk intends to give a brief overview of testthat and its internals, followed by a more detailed discussion about testwhat and the elegant adaptations that have been made to leverage testthat's functionality for an entirely different application.

Keywords: Testing, Learning R, OOP, testthat package, methods package

Failing fast and early: assertive/defensive programming for R data analysis pipelines

Tony Fischetti

College Factual

http://www.collegefactual.com

Abstract: In data analysis workflows that depend on un-sanitized data sets from external sources, it's very common that errors in data bring an analysis to a screeching halt. Oftentimes, these errors occur late in the analysis and provide no clear indication of which datum caused the error.

On occasion, the error resulting from bad data won't even appear to be a data error at all. Still worse, errors in data will pass through analysis without error, remain undetected, and produce inaccurate results.

The solution to the problem is to provide as much information as you can about how you expect the data to look up front so that any deviation from this expectation can be dealt with immediately.

We will talk about using the assertr package which supplies a suite of functions designed to verify assumptions about data early in analysis pipelines so that data errors are spotted early and can be addressed quickly.

Keywords: error-checking, assert, pipelines, data-checking, data-sanitizing

Getting your data into R

Hadley Wickham

Chief Scientist, RStudio

http://rstudio.com

Abstract: You can't use R for data analysis unless you can get your data into R. Getting your data into R can be a major hassle, so in the last few months I've been working hard to make it easier. I'll discuss the places you most often find data (databases, excel, text files, other statistical packages, web apis, and web pages) and the packages (DBI, xml2, jsonlite, haven, readr, exell) that make it easy to get your data into R.

Keywords: databases, csv, xml, API, SAS

A better way to manage hierarchical data

Christoph Glur

http://www.ipub.ch

Abstract: Data management is a huge time killer. This is especially true if you are working with data that is inherently non-tabular. The forthcoming data.tree package lets you organise hierarchical data in tree structures, and apply functions to the data by traversing the tree. In this presentation, I will cover the following subjects:

1. what is hierarchical data, and in what areas is it used (decision trees, finance, computer science, etc.)
2. what is reference semantics, and why is it useful when building trees / quick intro to reference classes and R6
3. features of the data.tree package
    a) building data.trees from scratch
    b) converting to and from data.frame
    c) tree traversal
4. applications
    a) ID3
    b) AHP (Analytic Hierarchy Process)
    c) TAA (finance, tactical asset allocation)

Keywords: Decision Trees, Hierarchic Data, Data Management, Reference Semantics

A proposal for distributed data-structures in R

Indrajit Roy, Michael Lawrence

HP Labs (US) and Genentech (US)

http://www.hpl.hp.com/

Abstract: Data sizes continue to increase, while single core performance has stagnated. To scale our computations, we need to distribute datasets across multiple machines. Thus, R needs standardized, idiomatic abstractions for computing on distributed data structures. R has many packages that provide parallelism constructs as well as bridges to distributed systems such as Hadoop. Unfortunately, each interface has its own syntax, parallelism techniques, and supported platform(s). As a consequence, contributors are forced to learn multiple idiosyncratic interfaces, and to restrict each implementation to a particular interface, thus limiting the applicability and adoption of their research and hampering interoperability.

Our proposal is to create a unified API for distributed computing. The API supports three shapes of data — lists, arrays and data frames— and enables the loading and basic manipulation of distributed data, including multiple modes of functional iteration (e.g., apply() like operations). In this talk we will discuss the proposed API, and how it can be implemented on top of existing distributed backends.

Keywords: Distributed computing, data-structures, big data, machine learning

Computational Performance

Chair: Dirk Eddelbuettel

Running R+Hadoop using Docker Containers

E. James Harner

West Virginia University

http://www.stat.wvu.edu/index.html

Abstract: There are numerous obstacles in implementing HDFS/Hadoop enabled to run R code. Scripts and packages must be distributed and possibly compiled. Just configuring Hadoop and R can be difficult for those not intimately familiar with the command line. This presentation will show how to easily start a multi-node Hadoop cluster using Docker and execute map-reduce jobs using R scripts. The cluster can be run both in a single VM on a laptop or on a cloud provider such as AWS or Azure. The implementation provides access to the Hadoop ecosystem built on HDFS/YARN, including the Spark, Storm, HBase, and Hive components. Multiple alternatives for writing R scripts and different deployment options are available. An experimental client front-end, which greatly simplifies interaction with HDFS/Hadoop, will be previewed.

Keywords: Hadoop, R scripts, Docker, Virtualization, Cloud computing

Algorithmic Differentiation for Extremum Estimation: An Introduction Using RcppEigen

Matt P. Dziubinski

Department of Mathematical Sciences, Aalborg University, Denmark
http://personprofil.aau.dk/profil/127800

Abstract: Algorithmic Differentiation (AD) enables automatic computation of the derivatives of a function implemented as a computer program. It's distinct from the numerical differentiation approximation methods (like finite differences), as it's exact to the maximum extent allowed by the machine precision. At the same time, it's free from the limitations of symbolic differentiation, since it works with actual computer programs (with branches, loops, allocations, and mutation) and not only pure, algebraic expressions.

AD is useful to a wide range of applications – in particular, fitting models via extremum estimators (i.e., estimators obtained as extrema of the given cost functions). This includes calibrating financial models, training machine learning models, or estimating statistical (or econometric) models. Numerical optimization algorithms necessary for model fitting benefit greatly from the high precision of the gradient obtained using AD – with a direct impact on the ultimate results' precision (more precise and stable parameter estimates, standard errors, confidence intervals).

This talk shows how to use AD from R – making the use of RcppEigen. We shall motivate the problem (including the issues with finite differences), introduce AD, and demonstrate its advantages over numerical approximations with a likelihood estimation example. We will end by speeding up the gradient computation via parallelization.

Keywords: algorithmic differentiation, numerical optimization, extremum estimation, maximum likelihood, RcppEigen

Improving computational performance with algorithm engineering

Kirill Müller

IVT, ETH Zurich

http://www.ivt.ethz.ch

Abstract: This talk presents three examples where algorithm engineering helps to achieve better runtime characteristics mostly with native R code. Weighted random sampling without replacement is solved in $O(nlogn)$ time by a novel reservoir sampling algorithm (instead of the straightforward $O(n^2)$ solution). Similar results are obtained for k-nearest-neighbor statistical matching, even if the Gower distance is used as a distance measure. For generalized raking—a survey calibration technique, a careful implementation of an existing algorithm is crucial to both correctness and computational performance. – The above examples are relevant, among others, for spatial microsimulation models and for generating synthetic populations for agent-based (e.g., transport) models. While the relatively weak performance of the standard implementations might go unnoticed for small examples, it quickly becomes an issue for larger real-world data sets. We evaluate correctness and runtime improvements for the new implementations.

Keywords: Algorithm engineering, Performance, Random sampling, Statistical matching, Survey calibration

Performance Analysis for Parallel R Programs: Towards Efficient Resource Utilization

Helena Kotthaus

Department of Computer Science 12, TU Dortmund University, Germany
http://ls12-www.cs.tu-dortmund.de/en/home/

Abstract: The R programming language is widely used in biostatistics with high-dimensional data sets. Here, a vast amount of resources is needed. Our tool traceR allows the user to profile the resource usage of an R application to locate bottlenecks and develop new optimizations.

Parallel computing is becoming a more and more popular option to reduce the effective runtime of compute-bound R applications. We therefore have improved traceR to allow for profiling parallel applications also. In contrast to existing profiling tools such as Rprof, traceR is directly integrated with the R interpreter. This enables the generation of more detailed and accurate data about memory and runtime behavior of an R application. Since the gain from parallel execution can be negated if the memory requirements of all parallel processes exceed the capacity of the system, this data can serve as a constraint to determine the maximum amount of parallelization.

For future work we want to use the profiling data produced by traceR to develop an optimized scheduling strategy for efficient resource utilization for parallel R programs. In this talk we will present our profiling tool traceR and how to apply it to analyze parallel R programs.

Keywords: Profiling, Performance Analyses, Tools, Distributed Computing, Machine Learning

Refactoring the xtable Package

David Scott

Department of Statistics, University of Auckland, Auckland, New Zealand
https://www.stat.auckland.ac.nz/

Abstract: The xtable package is a widely used package for including tables produced from R output in LaTeX and HTML documents. Unfortunately, a major function in the package, print.xtable has become unmanageably large over time. It now has 32 arguments and runs to 672 lines of R code. This project involved analysis of the code in print.xtable and refactoring using functions of much more moderate length. Test code was used to ensure that the refactored code produced the same results as the original code.

Keywords: Reproducible research, LaTeX, HTML

Kaleidoscope 3

Chair: Peter Dalgaard

Coding for the enterprise server - what does it mean for you?

Friedrich Schuster

HMS Analytical Software GmbH, Heidelberg, Germany

http://www.analytical-software.eu/de/

Abstract: For a data scientist, working with R typically means analyzing data for a particular task in a scientific context. As an R developer you will also be familiar with package development and associated tools and techniques.

Today R is also used for mission critical applications in production environments of large organizations. In this context an R developer faces a different, sometimes difficult situation with new challenges.

Some keywords concerning technical requirements are: reliability and high availability, scalability, compatibility and interoperability, security, maintainability. Other (non-technical) concerns are important as well: economic aspects (e.g. total cost of ownership), organizational aspects (different roles and responsibilities, collaboration in larger teams), and even legal requirements.

This presentation gives an overview of the factors that are important for the work of a developer in a corporate environment. It also tries to give some technical and non-technical advice for developing and maintaining client-server solutions with R.

Keywords: Enterprise, Server, Application, Development, Challenges

R as a citizen in a polyglot world - the promise of the Truffle framework

Lukas Stadler

Oracle Labs

https://labs.oracle.com/pls/apex/f?p=labs:bio:0:1917

Abstract: The days in which a problem could be solved within the confines of only one language are long gone - today's applications, including those written within the R language ecosystem, are multilingual, and today's programmers are polyglot.

This talk presents our vision of a multilingual environment centered around the R programming language. This environment uses the Truffle framework, backed by the Graal optimizing compiler, to provide an architecture that realizes all the advantages of polyglot programming, while at the same time offering solutions to hard questions like security, performance and debugging.

In the talk, we will introduce the architecture of our system, which is the foundation for a growing list of high-performance language implementations. We will also describe our progress towards utilizing this architecture to build a new high-performance compliant and just-in-time compiled R implementation and discuss how users can take advantage of interoperability with other languages supported by our system (e.g. JavaScript or Ruby).

Keywords: interoperability, polyglot, JavaScript, Ruby, compilation

Architect. An IDE for Data Science (and R)

Tobias Verbeke

Open Analytics NV

http://www.openanalytics.eu

Abstract: The life of a data scientist is a perilous adventure requiring many skills and good nerves. On Monday you write a Spark algorithm in Scala, on Tuesday you spin a Java image processing task in the cloud. Wednesday a Postgres DB needs your attention and Thursday you tweek an interactive visualization in Python. Finally on Friday you give your R packages some love. How to make sure you reach the weekend without getting insane? You open Architect on Monday morning and close it Friday evening.

Keywords: data science, integrated development environment, R

Distributed computing with R

Balasubramanian Narasimhan

Health Research and Policy & Statistics, Stanford University, Stanford, CA, USA
http://hrp.stanford.edu,http://statistics.stanford.edu

Abstract: Bringing together the information latent in distributed medical databases promises to personalize medical care by enabling reliable, stable modeling of outcomes with rich feature sets (including patient characteristics and treatments received). However, there are barriers to aggregation of medical data, due to lack of standardization of ontologies, privacy concerns, proprietary attitudes toward data, and a reluctance to give up control over end use. Statisticians have long known that aggregation of data is not always necessary for model fitting. In models based on maximizing a likelihood, the computations can be distributed, with aggregation limited to the intermediate results of calculations on local data, rather than raw data. We describe an R package distcomp to enable such computations and present several examples.

Keywords: distributed computing, cox regression, singular value decomposition

Business

Chair: Esben Høg

Statistical consulting using R: a DRY approach from the Australian outback.

Peter Baker

School of Public Health, University of Queensland, Herston, QLD, Australia

http://researchers.uq.edu.au/researcher/2181

Abstract: As a statistical consultant, I often find myself repeating the same steps when analysing data for different projects. Reusing R syntax and functions helps to improve efficiency and save time. Even bigger gains can be made by employing more automatic computing tools like GNU make and R packages like dryworkflow and ProjectTemplate. These tools help to implement a don't repeat yourself (DRY) approach.

Since the early 90s I've used GNU make to project manage data analysis using GENSTAT, SAS, R and other statistical packages. It is very efficient in only re–running analyses or producing reports when dependencies such as data, R syntax or Rmarkdown (.R, .Rnw or .Rmd) files change. Unfortunately, make doesn't provide rules for R and Rmarkdown so I have developed pattern rules which can be easily used by anyone. These make rules are also integrated into the R package 'dryworkflow' which follows the ideas in Long (2009) [The Workflow of Data Analysis Using Stata]. 'dryworkflow' streamlines many of the steps in a data analysis project.

Make pattern rules and the dryworkflow package will be briefly outlined and are available at github. Details may be found at http://www.petebaker.id.au/.

Keywords: data analysis, workflow, make, don't repeat yourself, reproducible research

Using R in Production

Stefan Milton Bache

Danske Commodities A/S, Aarhus, Denmark

http://www.danskecommodities.com

Abstract: The R programming language is becoming an integral part of data solutions and analyses in many production environments. This is no surprise: it is very powerful and goals are often very quickly accomplished with R compared to many of its competitors, and it is therefore a primary tool of choice for many statisticians, data scientists, and the like.

There can be some fundamental differences in reasonable requirements for R solutions produced for an academic purpose and those meant for a production environment. There are also some differences in the way they can *create value* to a business.

In this talk I share some of my thoughts and experiences with R in a production environment, and share some simple tips to keep in mind when preparing R solutions/scripts for business use, and demonstrate a few simple packages that can help avoid some common dangers and frustrations.

Keywords: Production environment, Rticularity, Code value

Hedging and Risk Management of CDOs portfolio with R

Giuseppe Bruno

Economics & statistics Dept, Bank of Italy

Abstract: Financial Institutions' portfolios feature a large number of credit risk instruments such as Collateralised Debt Obligations (CDO). Quite often the books of these institutions are composed of hundreds of CDO tranches and many more positions in credit risk flow instruments such as credit index and Credit Defaults Swap (CDS). These instruments provide the very basic weapons for the institution's hedging and speculative purposes. Performance analysis of these portfolios extends the analytical tools employed for the valuation and risk management of simple position on instruments such as CDO. The first goal of this work is to describe a Monte Carlo technique for computing the value of multiple tranches of synthetic CDOs built from a pool of hundreds of obligors. The Monte Carlo technique is then translated into R functions. The second goal is to introduce some R functions addressing the task of computing markto- market sensitivities with respect to spread, correlation etc. The final goal of the paper is to compute the optimal composition of an hedging tranche for an already given portfolio. For the optimization we compare the following three stochastic optimization:

1. Differential Evolution,
2. Genetic Algorithms,
3. Simulation Annealing.

Keywords: Collateral Debt obligations, Portfolio Hedging, Stochastic optimization, Genetic algorithms

Data Driven Customer Segmentation with R

Jim Porzak

DS4CI, El Cerrito, CA, USA

http://ds4ci.org

Abstract: To make strategic and tactical decisions businesses need to understand their customers. Simple, easy-to-understand, slowly varying segments are best for strategic decisions. Tactical marketers, on the other hand, need real time actionable segments. In general, the resulting segments need to make sense to the business user whether in senior management, marketing, sales, or product management.

We will cover four case studies showing how customer segmentation is done to satisfy strategic and tactical requirements. Specific R packages are referenced and simplified R code is shown.

The case studies are

1. Tenure based segments – based on a variation of survival analysis,
2. Recency/Frequency/Monetization (RMF) based segments using simple visualizations,
3. Cluster based segments using the flexclust package,
4. Life stage based segments using the TraMineR package.

Keywords: customer, marketing, segmentation, clustering

Spatial

Chair: Adrian Baddeley

Bringing Geospatial Tasks into the Mainstream of Business Analytics

Ian Cook

TIBCO Software Inc.

http://www.tibco.com/

Abstract: Businesses capture vast amounts of geospatial and location data. But historically, most business data analysts have lacked the tools and knowledge necessary to perform even basic manipulation and analysis of geospatial data. Traditionally, working with geospatial data required geographic information systems (GIS) software and specialists. In recent years, this has changed. Many R packages for working with geospatial data are available, and R has been integrated with mainstream business analytics software, enabling users to perform R-based tasks without writing R code. As a result, many geospatial data manipulation and analysis tasks are within the reach of business data analysts, and businesses can derive value from geospatial and location data with reduced complexity and cost.

Several basic geospatial data tasks will be demonstrated using TIBCO Spotfire and TIBCO Enterprise Runtime for R (TERR), including transformation of coordinate reference systems, spatial overlay, and calculation of geographical distances and areas. The R packages `sp`, `rgdal`, `geosphere`, and `wkb` will be used. Advanced geospatial analysis tasks will be briefly mentioned.

Keywords: geospatial, spatial, business, TERR

Novel hybrid spatial predictive methods of machine learning and geostatistics with applications to terrestrial and marine environments in Australia

Jin Li

Environmental Geoscience Division, Geoscience Australia, Canberra, Australia
http://www.ga.gov.au

Abstract: The accuracy of spatially continuous environmental data, usually generated from point samples using spatial prediction methods (SPMs), is crucial for evidence-informed environmental management and conservation. Improving the accuracy by identifying the most accurate methods is essential, but also challenging since the accuracy is often data specific and affected by multiple factors. Because of the high predictive accuracy of machine learning methods, especially random forest, they were introduced into spatial statistics by combining them with existing SPMs, which resulted in new hybrid methods with improved accuracy. This development opened an alternative source of methods for spatial prediction. In this study, we introduced these hybrid methods, along with the modelling procedure adopted to develop the final optimal predictive models. These methods were compared with the commonly used SPMs in R using cross-validation techniques based on both marine and terrestrial environmental data. We also addressed the following questions: 1) whether they are data-specific for marine environmental data, 2) whether input predictors affect their performance, and 3) whether they are equally applicable to terrestrial environmental data? This study provides suggestions and guidelines for the application of these hybrid methods to spatial predictive modelling not only in environmental sciences, but also in other relevant disciplines.

Keywords: spatial predictive methods, machine learning, geostatistics, spatial prediction, environmental modelling

Graphical Modelling of Multivariate Spatial Point Patterns

Matthias Eckardt

Department of Computer Science, Humboldt-Universität zu Berlin, Germany
https://www.informatik.hu-berlin.de/

Abstract: Multivariate spatial point patterns are of interest in many disciplines. Possible applications of multivariate point processes are manifold including spatial dependences of randomly distributed crimes, diseases or plants. Most commonly, spatial point patterns are analyzed using distance-based approaches including nearest-neighbor estimates or correlation functions. Although only marginally applied to spatial point processes spectral analysis presents a complementary approach with regard to the analysis of spatial point pattern. As advantage, related to periodicities of spatial structures spectral analysis provides a nonparametric approach without any prior structural assumption. Both approaches inherit limitations in case of analyzing data generated by multivariate point processes such as a strong increase of complexity or visual constraints beyond three dimensions. We address this problem of highly structured resp. highly complex multivariate point processes and propose a novel graphical model which is capturing the dependence structure of different events that occur randomly in space. We introduce a spatial structure graph which we recently implemented in R. Here, the edge set is identified by using the conditional spectral measures. Thereby, nodes are related to the components of a multivariate point process and edges express orthogonality relation in-between the single components. Examples will be taken from disease.

Keywords: Graphical Model, Spatial Point Pattern, Spectral Measures, High Dimensional, Structural Dependence

Spatial Econometrics Models with R-INLA

Virgilio Gomez-Rubio

Department of Mathematics, University of Castilla-La Mancha, Albacete, Spain
http://www.uclm.es

Abstract: The Integrated Nested Laplace Approximation (INLA) is widely used for approximate inference for Bayesian models. INLA focuses on marginal inference for models with observations from the Exponential family and latent effects that can be expressed as a Gaussian Markov Random Field. Recent developments include a new latent class to fit Spatial Econometrics models. This new 'slm' class is included in the R-INLA package and can be used to fit spatially autoregressive models that are either autoregressive on the response or the error term. By using the R-INLA package it is possible to use other built-in features such as computation of the DIC for model selection, prediction of missing observations and others. Another important topic in Spatial Econometrics is the computation of the impacts or spillover effects of the covariates. This can seldom be addressed with INLA as it requires multivariate inference, but we will also show how impacts can be approximated. The use of the 'slm' class will be illustrated using examples on housing in Boston and the probability of reopening a business in the aftermath of hurricane Katrina.

Keywords: Spatial Statistics, Spatial Econometrics, Integrated Nested Laplace Approximation, INLA

Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance

Sebastian Meyer

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

http://www.ebpi.uzh.ch/

Abstract: The availability of geocoded health data and the inherent temporal structure of communicable diseases have led to an increased interest in statistical models and software for spatio-temporal data with epidemic features. The open source R package surveillance can handle various levels of aggregation at which infective events have been recorded: individual-level time-stamped georeferenced data (case reports) in either continuous space or discrete space, as well as counts aggregated by period and region. For each of these data types, the surveillance package implements tools for visualization, likelihoood inference and simulation from recently developed statistical regression frameworks capturing endemic and epidemic dynamics. This presentation is intended as a guide to the spatio-temporal modeling of epidemic phenomena, exemplified by analyses of public health surveillance data on measles and invasive meningococcal disease.

Keywords: spatio-temporal endemic-epidemic modeling, infectious disease epidemiology, self- exciting point processes, multivariate time series of counts

Databases

Chair: Thomas Lumley

Rango - Databases made easy

Willem Ligtenberg

Open Analytics N.V., Antwerpen, Belgium

http://www.openanalytics.eu/

Abstract: Rango is a package that makes it easy to use relational databases directly from R and without a line of SQL (an Object Relational Mapper in techspeak).

Let's consider an example. Assume we have a database of music tracks. This simple version consists of two tables, one containing information about the artists (artist) and one containing information about the tracks (track).

Rango allows you to create new artist or track objects in R and then either store or retrieve them from the database. The relationship between a track and an artist is stored as an attribute in the track object. This allows Rango to automatically create joins when you want to make a selection in one table based upon another table. In our example, we could ask for all tracks by a given artist.

Rango optimizes where possible. For example, we use lazy loading to only retrieve information from associated tables upon access. Rango can cache results of queries making subsequent retrievals faster. Comparison operators have also been implemented to easily retrieve lists of objects.

Lastly, all functions can be used with the same syntax regardless of the database backend. Currently supported database backends are SQLite and PostgreSQL.

Keywords: PostgreSQL, SQLite, ORM, Object Relational Mapper, database

Ad-Hoc User-Defined Functions for MonetDB with R

Hannes Mühleisen

Database Architectures Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

https://www.cwi.nl/research-groups/database-architectures

Abstract: Various capable DBI connectors allow R to retrieve data from all major relational databases. Unfortunately, the overhead associated with a plethora of data conversions on the way makes this approach unfeasible as the amount relevant to the analysis grows. R's support for running as a library allows for a different approach, where a in-database User-Defined Function (UDF) is implemented in R, which is executed as relational operator at query runtime. Here, transfer overhead is greatly reduced, which improves performance. Packages such as PL/R for PostgreSQL have been available for years [1]. However, there is still considerable conversion overhead, since traditional databases use the Row-major format for storing data, whereas R is most efficient on columnar data. We have embedded the R environment into MonetDB [2], an Open Source columnar relational database. The almost identical in-memory representations of MonetDB tables and R data.frame objects allows an efficient handover of data between environments. In our experiments, we were able to outperform PL/R by several orders of magnitude. In out talk, I will describe how the extension can be used to best leverage the power of both worlds, highly optimized relational operators together with highly specialised statistical methods.

[1] http://www.joeconway.com/plr/

[2] https://www.monetdb.org/content/embedded-r-monetdb

Keywords: Column Store, Database, Bigger datasets

R database connectivity: what did we leave behind?

Mateusz Żółtak

Educational Research Institute, Warsaw, Poland

http://ibe.edu.pl

Abstract: The idea of having one common set of functions (API) to connect to any relational database system arose very early in the R community. The idea was implemented as the DBI package, released in December 2001. Unfortunately, since the beginning the DBI package has focused on translating R's world into relational databases' world rather than the opposite. This approach has resulted in lacking guidelines on implementing various common and important features of relational databases, e. g. schemes, views, parameterized statements, etc. which do not have any equivalents in R's world. Consequently, R users have inconsistently implemented such features in various relational database connectivity packages, stifling the initial idea of one common API. Nowadays, developing a common API is more important than ever for at least two reasons. First, R is more and more frequently used in business solutions where the most common way of storing data are big and complicated relational databases. Second, following the big data revolution, relational database systems are becoming more and more popular with researchers using R. In my talk, I would like to analyse the problem and its consequences in more detail, and discuss possible solutions.

Keywords: relational databases, common API, DBI

jsonlite and mongolite

Jeroen Ooms

Department of Statistics, UCLA, Los Angeles
http://statistics.ucla.edu/

Abstract: The jsonlite package provides a powerful JSON parser and generator that has become one of standard methods for getting data in and out of R. We discuss some recent additions to the package, in particular support streaming (large) data over http(s) connections. We then introduce the new mongolite package: a high-performance MongoDB client based on jsonlite. MongoDB (from "humongous") is a popular open-source document database for storing and manipulating very big JSON structures. It includes a JSON query language and an embedded V8 engine for in-database aggregation and map-reduce. We show how mongolite makes inserting and retrieving R data to/from a database as easy as converting it to/from JSON, without the bureaucracy that comes with traditional databases. Users that are already familiar with the JSON format might find MongoDB a great companion to the R language and will enjoy the benefits of using a single format for both serialization and persistency of data.

Keywords: json, database, web, interoperability

Using R Efficiently with Large Databases

Michael Wurst

IBM Research and Development, Germany

Abstract: This session gives a survey on open technologies that help statisticians to work more efficiently with large data warehousing systems. We will especially discuss how data warehouse scale-out and indexing structures, such as columnar storage, can be exploited from R to increase scalability and reduce response time. The session should also serve as a platform to discuss limitations, open challenges and feature requests from the point of view of R users that work with large databases.

Keywords: Databases, Warehousing, SQL, Scalability

Kaleidoscope 4

Chair: Susan Holmes

While my base R gently weeps

A. Jonathan R. Godfrey

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

http://massey.ac.nz

Abstract: Blind users want the same from base R as does any other user. More importantly though, we also want the same benefits that front-ends and integrated development environments offer our classmates and colleagues. My assessment of all of the excellent tools tested to date is that none will meet the needs of blind users in the short to medium term. The BrailleR package hopes to fill this gap.

While the BrailleR package's original intention was to provide text representations of graphics and some convenience functions for helping novice users get started, other (more practical) issues have been given attention over the last twelve months. The package is receiving much-needed time and energy since unleashing the benefits of Rmarkdown via the knitr package.

Increased interaction with blind users via workshops and an email list (called "BlindRUG") have shown the difficulties novice blind users are facing when getting started. Recent developments have included: a simple text editor for processing Rmarkdown files; functions to help set up the user's R installation; a method to view and edit datasets by linking to standard spreadsheet software; and, alternative ways of processing R scripts and Rmd documents. The talk will include practical demonstrations of these features.

Keywords: blindness, braille, screen reader, reproducible research

Rapid Deployment of Automatic Scoring Models to Hadoop Production Systems

Amitai Golub

Innogames GmbH, Hamburg, Germany

http://corporate.innogames.com

Abstract: Predictive modelling in R is the bread and butter of the data scientist, but deploying these analyses into production systems often runs into many barriers. These difficulties stem mostly from issues pertaining to scalability and communication between different systems, very commonly R & Hadoop.

A classic example in the gaming industry is player retention, often critical for the success of a company, but notoriously hard to predict. Once a method has been found to work, we would like to deploy quickly in order to scale to as many players as possible.

We show a three step process allowing data scientists to quickly develop models and deploy them to Hadoop systems, using the RHive, rmr2, and plyrmr packages. This process yields a fully automatic churn prediction model run automatically by a cron job. Furthermore, once initial deployment is done, further user scoring models can be easily added to the system.

Keywords: Automation, MapReduce, Big Data , RHadoop, RHive

Fast, stable and scalable true radix sorting

Matt Dowle

H2O.ai, California

http://h2o.ai

Abstract: This talk will present the details and benchmarks of the fast and stable radix sort implementation in data.table::forder. For example on 500 million random numerics (4 GB), base R takes approximately 22 minutes vs forder at 2 minutes. The pros and cons of most-significant-digit (forwards) and least-significant-digit (backwards) will be discussed as well as application to all types: integer with large range (>1e5), numeric and character. We hope to find a sponsor from the R core team to help us include this method in base R where it could benefit the community automatically. Package data.table will only be mentioned in passing. The work builds on articles by Terdiman, 2000 and Herf, 2001 and is joint work with Arun Srinivasan.

Keywords: fast, sorting, large, data

Fast, flexible and memory efficient data manipulation using data.table

Arunkumar Srinivasan

Open Analytics

http://www.openanalytics.eu

Abstract: Data manipulation operations such as subset, join, aggregation, update etc. are all inherently related. By keeping these related operations together, the data.table package allows for fast and memory efficient data analysis of large data (1GB – 100GB or more in RAM). Although speed benefits are pronounced on larger datasets, many also use it on small data for its concise and consistent syntax.

Data.table inherits from and extends from data.frame. The general form of its syntax including chaining is:

```
DT[where, select | update, group by][order by][...][...]
```

This talk showcases several recent enhancements and features implemented in data.table such as:

- Efficiently reading file(s) (rbindlist + lapply + fread)
- Quick review of reasons behind speed and memory efficiency
- Aggregations and updates during joins (by = .EACHI)
- Automatic indexing based subsets (using binary search)
- Efficient reshaping of multiple columns simultaneously (melt + dcast)
- Fast joining over intervals, e.g., genomic data, date ranges (foverlaps)
- Fast and memory efficient ordering by reference (setorder)

with clear examples and benchmarks on relatively large datasets (1GB to 3GB).

References:

M Dowle, T Short, S Lianoglou, A Srinivasan, R Saporta, E Antonyan (2008-2015). Data.table: Extension of data.frame. https://github.com/Rdatatable/data.table

On CRAN: http://cran.r-project.org/web/packages/data.table/index.html

Keywords: Large data, Auto-indexing, aggregations and reshaping, overlapping joins, update by reference

Medicine

Chair: Heather Turner

Phenotypic deconvolution: the next frontier in pharma

Marvin Steijaert[†], ***Vladimir Chupakhin***[‡], ***Hugo Ceulemans***[‡], ***Joerg Wegner***[‡]

[†]*Open Analytics NV* [‡]*Janssen Pharmaceutica*

http://www.openanalytics.eu/

Abstract: Pharmaceutical drug design faces high failure rates in clinical trials. One strategy to reduced this failure rate is the use of disease-relevant phenotypic assays. These assays measure compound activity in a multi-target cellular context, while classic assays aim at isolated targets. Drawback of phenotypic assays is the lack of knowledge about modes of action of active compounds (i.e., involved proteins and potential drug targets). This knowledge is crucial for further development towards a potent drug with little side effects. We use machine learning methods to predict the most likely modes of action.

Our (R and C++) modeling pipeline uses a two-layer approach. The first layer predicts compound activity on single protein targets at five different concentrations. As input we use the measured activity of 1.5M compounds on 1.5k targets. Due to the high costs of generating such amounts of data, only 2% of the compound-target combinations have been measured. Hence the need for predictive modeling. The second layer combines these predictions and selects the features (targets) that can best explain the phenotypic assay data.

Keywords: machine learning, feature selection, bioassays, chemical fingerprints

medplot: A Web Application for Dynamic Summary and Analysis of Longitudinal Medical Data Based on R and shiny

Lara Lusa

Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia

http://ibmi.mf.uni-lj.si/en

Abstract: We present medplot, an interactive web application that simplifies the exploration and analysis of longitudinal data. The web application was developed using the framework offered by the shiny R package, which considerably simplifies the creation of web applications based on code written in R. Longitudinal data arise often in biomedical studies, where patients are often evaluated numerous times and a large number of variables are recorded at each time-point. Researcher with biomedical background often find difficulties in the use of specialized statistical software, which offers the capability of correctly and flexibly analyze this type of data. medplot can be used to summarize, visualize and analyze data by researchers that are not familiar with statistical programs. The summary tools produce publication-ready tables and graphs. The analysis tools include features that are seldom available in spreadsheet software, such as correction for multiple testing, repeated measurement analyses and flexible non-linear modeling of the association of the numerical variables with the outcome. medplot is freely available and open source, it has an intuitive graphical user interface (GUI), it is accessible via the Internet and can be used within a web browser, without the need for installing and maintaining programs locally on the user's computer.

Keywords: web application, shiny, longitudinal data

Using R and free software to improve the delivery of life changing medicine to patients

David Ruau / Paul Metcalfe

AstraZeneca, Cambridge, UK

http://www.astrazeneca.com

Abstract: The methods used by statisticians in the pharmaceutical industry rapidly evolve through constant challenges, but the toolset available to statisticians and data scientists has been largely static. However, a new generation of quantitative scientists is introducing tools like R and Shiny. In this presentation we will present some use cases where the rapid development and deployment of R packages has made a huge difference. In the first example, we implemented a classic Bayesian method used to predict the probability of technical success in phase III of clinical trials based on results from phase II. The release of this R package (assurance) first internally and then on GitHub not only enabled project teams to consistently integrate this information when designing their trial, but also enabled us to validate the R package through a wider community. In the second use case we looked into one of the challenges our statisticians have when interacting with contract research organisations. Through development of an R package and a Shiny interface we harmonized the simulation and prediction of time-to-event in oncology clinical trials saving considerable time and improving consistency.

Keywords: Clinical trials, bayesian inference, sample size, power, assurance

Stratified medicine using the partykit package

Heidi Seibold

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

http://www.ebpi.uzh.ch/

Abstract: The identification of patient subgroups with differential treatment effects is the first step towards individualised treatments. A current draft guideline by the EMA discusses potentials and problems in subgroup analyses and formulated challenges to the development of appropriate statistical procedures for the data-driven identification of patient subgroups. We introduce model-based recursive partitioning – which is implemented in the R partykit package – as a procedure for the automated detection of patient subgroups that are identifiable by predictive factors. The method starts with a model for the overall treatment effect as defined for the primary analysis in the study protocol and uses measures for detecting parameter instabilities in this treatment effect. The procedure produces a segmented model with differential treatment parameters corresponding to each patient subgroup. The subgroups are linked to predictive factors by means of a decision tree. The method is applied to the search for subgroups of patients suffering from amyotrophic lateral sclerosis that differ with respect to their Riluzole treatment effect, the only currently approved drug for this disease.

Keywords: partykit, subgroup analysis, stratified medicine, trees, parametric models

Regression

Chair: Rasmus Waagepetersen

The ilc package

Han Lin Shang

Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University, Canberra, Australia

http://www.rsfas.anu.edu.au

Abstract: We implement a specialised iterative regression methodology in R for the analysis of age-period mortality data based on a class of generalised Lee-Carter (LC) type modelling structures. The LC-based modelling frameworks is viewed in the current literature as among the most efficient and transparent methods of modelling and forecasting mortality improvements. Thus, we make use of the Generalised Linear Model (GLM) modelling approach discussed in Renshaw and Haberman (2006), which extends the basic LC model and proposes to make use of a tailored iterative process to generate parameter estimates based on Poisson likelihood. Furthermore, building on this methodology we develop and implement a stratified LC model for the measurement of the additive effect on the log scale of an explanatory factor (other than age and time). This modelling methodology is implemented in a publicly available ilc package that facilitates both the preparation of mortality data and the fitting and analysis of the given log-linear modelling structures. The package also incorporates methods to produce forecasts of future mortality rates and to compute the corresponding future life expectancy.

Keywords: generalised/extended Lee-Carter models, age-period-cohort models, iterative estimation approach

Approximately Exact Calculations for Linear Mixed Models

Andrew Bray

Mathematics Department, Reed College, Portland, Oregon, USA

http://academic.reed.edu/math/

Abstract: One of the primary ways that statisticians learn about a linear mixed model in the context of a given data set is to consider the restricted likelihood (RL) function or the joint posterior. Of particular interest are the regions of the parameter space where these functions are high. These regions are traditionally found using general optimizing algorithms that may fail to find the global optimum. In this talk we present a method to optimize the RL function and posterior that is specific to mixed models with two variances. It is a branch and bound algorithm that evaluates the function to an arbitrary degree of precision within a given region of the parameter space, ensuring that no optima are excluded.

The algorithm has been implemented in R with attention paid to making the computing time comparable to the existing algorithms used in the popular lme4 package. We will discuss the approaches to implementation that we have tried, and compare their resulting performance.

Keywords: linear models, mixed models, optimization, data.table, likelihood

Shiny application for analyzing consumer preference and sensory data in a mixed effects model framework: introducing SensMixed package

Alexandra Kuznetsova

DTU Compute

http://www.dtu.dk/

Abstract: Often too simplistic mixed effects models are used in analysis of sensory studies. One of the reasons to that is connected with the fact that no open-source software or application is available that is specifically dedicated to analyze these types of studies and can handle complex mixed effects models.

We present the SensMixed package, that offers analysis of sensory and consumer data within a mixed effects model framework (A. Kuznetsova et al. 2015). The package provides a number of options for the model building of the mixed effects models, which are constructed with the lmer function of the lme4 package (Bates et al., 2014) and are tested with the lmerTest (A. Kuznetsova et al., 2013) package.

The SensMixed package includes a shiny application (RStudio, 2013), which apart from providing the GUI for the analysis of mixed effects models, also includes such crucial functionalities as importing the data in different formats, presenting results in tables and plots as well as saving them. This makes the package together with the application very valuable for non-statisticians. The usefulness of the package will be illustrated on examples coming from the sensory and consumer data.

The SensMixed package can be download from https://r-forge.r-project.org/R/?group_id=1433

Keywords: consumer and sensory data, mixed effects models, shiny

Spatial regression of quantiles based on parametric distributions

Chenjerai Kathy Mutambanengwe

OpenAnalytics, Antwerp, Belgium

http://www.openanalytics.eu

Abstract: Quantile regression has garnered great interest in the recent years. This is because the mean (or variance) of a random variable is not always of interest to investigators. For the special case of spatially correlated data, estimation of the quantiles becomes slightly more challenging due to the need to take into account the spatial correlations between measurements, and the most commonly used R package quantreq does not exhaustively cater for such situations. We develop Bayesian methods for estimating the spatially varying quantiles, whilst allowing the effects of covariates to also vary conditionally on the quantiles. In this presentation we focus on the log-location-scale family of distributions, thereby restricting to a model-based approach as opposed to the model-free approach. These methods are made available in the spatQuantReg package and implemented in R. The methodology will be highlighted in the first part of the presentation, and some analysis will be made on a real data example of TSH levels from Galicia, Spain.

Keywords: quantile regression, spatial data, bayesian

glmmsr: fitting GLMMs with sequential reduction

Helen Ogden

Department of Statistics, University of Warwick, Coventry, UK

http://warwick.ac.uk/statistics

Abstract: Generalized linear mixed models (GLMMs) are an important and widely used model class. In R, we can fit these models with the lme4 package, but there are some limitations. First, except in very simple cases, lme4 uses a Laplace approximation to the likelihood for inference, which may be of poor quality in some cases. Second, it is difficult to fit some GLMMs, such as pairwise comparison models, with lme4. The glmmsr package offers progress on both of these problems. It implements the sequential reduction approximation to the likelihood, controlled by a parameter which allows the user to trade-off the accuracy of the approximation against the time taken to compute it. The interface of glmmsr is an extension of that of lme4, allowing easy fitting of pairwise comparison and many other interesting models.

Keywords: lme4, mixed model, intractable likelihood, pairwise comparison, BradleyTerry2

Commercial Offerings

Chair: Romain François

Supporting the "Rapi" C-language API in an R-compatible engine

Michael Sannella

TIBCO Software Inc., Seattle, WA 98109, USA

http://spotfire.tibco.com

Abstract: Much of the value of the R engine is provided by the large set of external packages available from repositories such as CRAN. While many of these packages are "pure" R-language packages, many important and popular packages contain C-language libraries that access the R engine through R's C-language API (which I call "Rapi").

My group at TIBCO is currently developing TERR (TIBCO Enterprise Runtime for R), an R-compatible engine. Our users want to be able to use many CRAN packages including those incorporating C libraries, so TERR has to support the Rapi API. Our goal is to allow our users to download binary packages from CRAN (or build source packages with open-source R), and then load and use them unchanged in TERR.

This presentation will discuss our experiences working with packages using Rapi, the current state of TERR's support for Rapi, and our efforts to increase the set of packages TERR can support.

Keywords: R packages, TERR, implementation, API

Enabling R for Big Data with PL/R and PivotalR: Real World Examples on Hadoop & MPP Databases

Woo J. Jung

Pivotal

http://www.pivotal.io

Abstract: Exporting data from Hadoop or a database and importing into a server or desktop environment with R is not an ideal workflow for big data analytics – among many issues, end users may face challenges around the memory/scalability limitations of R and the costs associated with transferring large amounts of data between different platforms. We explore a closer integration of R with big data platforms such as Hadoop and MPP databases (i.e. Pivotal HD, HAWQ, Greenplum Database), focusing on real world examples drawn from the work of data scientists at Pivotal. We describe how open source tools such as PL/R and PivotalR enable R for big data, and walk through case studies from industries such as retail, telco, and utilities that illustrate the flexibility & performance of these extensions. Further, we will also touch on how tools such as PL/R and PivotalR can be leveraged for scalable statistical & machine learning algorithm development on big data platforms, using Bayesian Hierarchical Regression with Gibbs Sampling as an illustrative example.

Keywords: big data, Hadoop, parallel, applications, PL/R, PivotalR

The DataRobot R Package

Ron Pearson

Data Scientist, *DataRobot*, *Boston*, *MA*, *USA*

http://www.datarobot.com/

Abstract: This talk describes DataRobot, a new R package that allows users to interact with the DataRobot modeling engine. This API allows the user to create a new DataRobot project by specifying a dataset, a target variable, and a fitting metric. The modeling engine then builds a large collection of models (typically 30 - 60), which are fit and tuned by cross-validation and ranked in order of their validation set performance. The models in this collection encompass a wide variety of types, including linear regression models, random forests, boosted trees, and support vector machines. The R package described here allows the user to characterize the models obtained, add custom R models to the project, and generate predictions from any of these models for new datasets. This talk demonstrates the DataRobot package by applying it to two examples. The first is a random permutation-based assessment of variable influence, obtaining results for all of the models in the project, which were fit to data from the mlbench.friedman1 simulator in the R package mlbench. The second example demonstrates the use of partial dependence plots to understand performance differences between very different model structures, all fit to a concrete compressive strength dataset.

Keywords: variable importance measures, partial dependence plots, model comparison, concrete compressive strength, mlbench.friedman1 simulator

Applying the R Language in Streaming Applications and Business Intelligence

Lou Bajuk-Yorgan

TIBCO Software

http://spotfire.tibco.com

Abstract: R provides tremendous value to statisticians and data scientists, however they are often challenged to integrate their work and extend that value to the rest of their organization. This presentation will demonstrate how the R language can be used in Business Intelligence applications (such as Financial Planning and Budgeting, Marketing Analysis, and Sales Forecasting) to put advanced analytics into the hands of a wider pool of decisions makers. We will also show how R can be used in streaming applications (such as TIBCO Streambase) to rapidly build, deploy and iterate predictive models for real-time decisions. TIBCO's enterprise platform for the R language, TIBCO Enterprise Runtime for R (TERR) will be discussed, and examples will include fraud detection, marketing upsell and predictive maintenance.

Keywords: Business, TERR, Business intelligence, Streaming data, Real time

Interactive graphics

Chair: Di Cook

D3 and R Shiny – Making your graphs come to life

Monika Huhn; Jesper Havsol; Daniel Goude; Martin Karpefors

Biometric & Information Sciences, AstraZeneca, Molndal, Sweden

http://www.astrazeneca.se/

Abstract: The human visual perception system has an extraordinarily capacity for graphical interpretation, and an effective visualization will therefore make complex data more accessible, understandable and usable. What makes the visualizations even more powerful is the addition of interactivity to the plots, which will empower less technical collaborators to explore otherwise hidden information. Furthermore, animating data over time can reveal important temporal patterns.

Currently available visualization tools often fall short of providing easy-to-use and visually attractive animation functionalities. In this talk, we show how the combination of interactivity, provided by the R Shiny package, and dynamics, provided by the D3.js javascript package, can bring a new dimension to the analyses.

D3 has unlimited flexibility and very good data transition capabilities and the R Shiny package enables interactive web applications straight from R. Consequently, connecting these two pieces with the excellent data manipulation properties in R, results in a powerful web application. The application was originally developed to enable clinicians to look at medical data and investigate changes over time, but could be useful in any field carrying time-changing information.

Keywords: Shiny, D3, Animation, Interactivity, Data Visualization

Interactive Graphics with ggplot2 and gridSVG

Michael Sachs

Biometric Research Branch, National Cancer Institute, Bethesda, Maryland, United States of America

http://brb.nci.nih.gov/

Abstract: Interactive statistical graphics can be useful in data analysis and reporting. They allow supplemental information to be displayed along with a standard chart, without diluting the main message. The package ggplot2 provides a powerful interface for creating high quality statistical graphics, while gridSVG converts grid graphics objects to svg objects that can be rendered in web browsers. We combine the two, along with a bit of JavaScript, to create interactive statistical graphics for use on the web. The idea is illustrated with examples for plotting receiver operating characteristic curves, and Kaplan-Meier survival curves, with interactive features bound to hover and click events. This approach differs from many others in that the figure rendering is handled by R, instead of a JavaScript library. JavaScript is used instead to bind interactive events to the svg objects. Our interface is based on the main strengths of R: the statistical computations and graphics rendering allowing for seamless transitions between static and interactive plots, retaining the R/ggplot2 style and allowing visual consistency across document types.

Keywords: graphics, interactive graphics, ggplot2, gridSVG

Interactive visualization using htmlwidgets and Shiny

Joe Cheng

RStudio, Inc.

http://www.rstudio.com/

Abstract: The htmlwidgets and Shiny packages are designed to bring interactive JavaScript data visualization technology to R. Used together or separately, these packages can be used to easily create exploratory data tools, interactive HTML reports, and web applications and dashboards.

This talk will demonstrate some interesting uses of Shiny, htmlwidgets, and related packages.

Keywords: visualization, javascript, shiny, htmlwidgets

Interactive Data Visualization using the Loon package

Adrian Waddell

University of Waterloo

https://uwaterloo.ca/statistics-and-actuarial-science/

Abstract: The loon R package provides an extendable and highly interactive data visualization toolkit for R users based on the open source Tcl/Tk package of the same name. Interactions with plots are provided with mouse and keyboard gestures as well as via command line control and with inspectors that provide graphical user interfaces (GUIs) for modifying and overseeing plots.

In this talk, we will demonstrate an exploratory visual data analysis using loon. In particular, we will analyze a dataset on the distribution of visible minority populations across Canada. Relevant features of loon used in this analysis include: zooming, panning, selection and modification of points, dynamic linking of plots, layering of visual information such as maps and regression lines, and custom point glyphs such as text or star glyphs.

We conclude this talk by giving an overview of loon's design. This will include a discussion of loon's extensive event bindings that can be used to extend and customize loon's behavior and features.

Loon's capabilities are very useful for statistical analyses such as interactive exploratory data analysis, sensitivity analysis, animation, teaching, and creating new graphical user interfaces.

Keywords: Interactive Data Visualization, Explorative Data Analysis, Graphical User Interfaces

New interactive visualization tools for exploring high dimensional data in R

Wayne Oldford

Department of Stats & Act. Sci., University of Waterloo, Canada
https://math.uwaterloo.ca

Abstract: Graphs whose nodes represent low dimensional data spaces and whose edges are transitions between them provide a relatively simple visual aid for navigating high dimensional data spaces via low dimensional trajectories (see Hurley and Oldford, Comput Stat 2011). The R package RnavGraph (Waddell and Oldford, useR! 2011) provided a proof of concept for this new type of interactive data visualization method in R.

In this presentation, we take advantage of the extendable and highly interactive R package loon to present new and integrated tools for exploring high dimensional data in R. Novel tools include, for example, various navigational graphs, interactive scagnostic scatterplot matrices, interactive serial axes displays (parallel and radial coordinates), and novel event bindings. Because loon's event bindings can evaluate callbacks to any R function, the navigation graphs can drive any display (or analysis) in R, including loon's own interactive plots.

Keywords: Data Visualisation, Interactive graphics, High dimensional data, Exploratory data analysis

Kaleidoscope 5

Chair: Søren Højsgaard

Formalising R Development - ValidR Enterprise

Aimee Gott

Mango Solutions

http://www.mango-solutions.com

Abstract: As R developers we all write code every day, but most of us are statisticians or data scientists. We haven't been trained to write code. We haven't leant about version control, unit testing or continuous integration. At Mango we have a development team who have been trained to do all of this and can teach us a lot about best practices in software development. This has allowed us to develop a rigid framework for development that satisfies regulators in a number of industries and has become the basis for our validation of R and consequently the framework of ValidR Enterprise.

In this talk we will look at some of the challenges faced when formally developing R code and how we have used ideas from computer science to build a formal development environment. We will also look at how this has been incorporated into ValidR and become the framework for ValidR Enterprise.

Keywords: development, regulation, validation

CXXR: Modernizing the R Interpreter

Karl Millar

Google

https://www.google.com

Abstract: CXXR is a project to refactor the internals of the current R interpreter using C++ and modern software engineering techniques.

The goal of the project is to produce a fully compatible, open source, production quality R implementation suitable for everyday use but with significantly better performance than current R implementations.

Many techniques that are used to achieve high performance in languages such as Python and JavaScript are also applicable to the R language, so similar performance is likely to be achievable. Additionally, the Riposte project has demonstrated that the loop fusion and automatic parallelization techniques of vector languages, such as APL, can be used to significantly accelerate and parallelize well-vectorized R code.

We demonstrate the progress that has been made towards these goals so far and discuss the roadmap to a fast, robust R implementation.

Keywords: CXXR, C++, Performance, LLVM

Fun times with R and Google Sheets

Jennifer Bryan

Statistics and the Michael Smith Labs, University of British Columbia, Vancouver, Canada

http://www.stat.ubc.ca

Abstract: Spreadsheets are something like the Ellis Island of R: lots of users and data come in through this gateway. And like Ellis Island, it is wonderful that this portal exists, but we can probably make the facilities more welcoming!

I've worked with Joanna Zhao to create the R package 'googlesheets', which allows the useR to work with public and private Google Sheets from R. 'googlesheets' wraps both the Sheets and Drive APIs, so the useR can consume data, edit data, and create, delete, rename, copy, upload, and download spreadsheets and worksheets.

Talk topics (I'll select and develop based on session context and time available):

- The potential for Google Sheets to act as a low-barrier CMS and webscraper, making it easier for novices to marshal data they want to analyze with R.
- The joys of test automation and continuous integration for an API wrapper package, especially with OAuth2.
- Using Google Sheets as a data store for a Shiny app.
- Is it lunacy to parse the formulas in spreadsheets and translate the computations into R code? Motivation: to help those trying to transition from spreadsheet users to R users.

Keywords: spreadsheets, API, package, testing, data ingest

A Comparative Study of Complex Estimation Software

Jonathan Digby-North

Survey Methodology and Statistical Computing, Office for National Statistics

http://www.ons.gov.uk/ons/guide-method/method-quality/index.html

Abstract: The Office for National Statistics (ONS) uses sample surveys to provide the UK with key social and economic statistics. For a number of these surveys, the calibration and weighting is implemented via Statistics Canada's SAS-based Generalized Estimation System (GES). This software, which must first be purchased, can only be used with an appropriate SAS licence.

As part of several ongoing initiatives to explore alternative open source solutions at ONS, this project assesses the feasibility of replacing GES with the free R package 'ReGenesees', written and developed by the Italian Statistics Office (Istat). A selection of household and business surveys – each of which currently uses GES for a different task or is unique in some way – was used to compare the two tools in terms of their functionality, performance and ease of use.

The overall conclusion of this project was that the 'ReGenesees' R package is a viable (and perhaps favourable) alternative to GES for the Office.

Keywords: Surveys, Calibration, Estimation, Software

Software Standards in the R Community: An Analysis

Oliver Keyes

Wikimedia Foundation

http://wikimediafoundation.org/wiki/Home

Abstract: This work reports on research by Oliver Keyes and Jennifer Bryan on the software engineering standards used by the R community in CRAN-hosted packages, along with ways to improve those standards.

Keywords: standards, software, testing, documentation, community

Teaching 1

Chair: Rasmus Waagepetersen

SWOT analysis on using R for online training

Miranda Y Mortlock

School of Agriculture and Food Science, University of Queensland, St Lucia, Qld, Australia

http://www.uq.edu.au/

Abstract: This will present a SWOT (Strengths Weaknesses Opportunities and Threats) analysis on using R in an online courses. R and RStudio were chosen as the software to extend statistical support via a bespoke online statistical training (BeST) site. The site is being developed as a Special Private Online Course (SPOC) and has potential to become a Massive Open Online Course (MOOC) in the future. It is aimed at use in Africa and South Asia, and will have the ability to influence scientist that are working in research projects supporting millions of small farmers in these regions. BeST is mostly aimed at African agricultural scientists who do not have statistical support because they may be either based regionally or are the field and away from consulting centres of statistics. The BeST site is under development, using a modular approach to support a range of applied methods. The use of R and RStudio will be given a critical but practical analysis.

Keywords: Online training, Special Private Online Course (SPOC) , applied statistics, Massive Open Online Course (MOOC) , agricultural science

Manipulation of Discrete Random Variables in R with discreteRV

Eric Hare

Department of Statistics, Iowa State University, USA

http://www.stat.iastate.edu

Abstract: A prominent issue in statistics education is the sometimes large disparity between the theoretical and the computational coursework. discreteRV is an R package for manipulation of discrete random variables which uses clean and familiar syntax similar to the mathematical notation in introductory probability courses. The package offers functions that are simple enough for users with little experience with statistical programming, but has advanced features which are suitable for a large number of more complex applications. In this talk, I will introduce and motivate discreteRV, describe its functionality, and provide reproducible examples illustrating its use.

Keywords: discrete, random, variables, probability, education

Teaching R in heterogeneous settings: Lessons learned

Matthias Gehrke

ifes, Institut fuer Empirie und Statistik, FOM University of Applied Sciences, Germany

http://www.fom-ifes.de

Abstract: There is a rich field of economic and business applications for statistical data analysis and R: From Human Resources to Finance. In 2013 the FOM, a private university of applied sciences for professionals studying while working with more than 30 study centres across Germany, decided to use R compulsory in all statistical courses in all the different Master programs and in all study centres.

By using the R Commander (Fox, 2005), real life data sets, step-by-step installation guidelines as well offering a portable version with preinstalled packages and a support email address we tried to facilitate and motivate the start and working with R.

Coping with the heterogeneity of students, lecturers and computer platforms like Windows and Mac OS there are some pitfalls in such a teaching project, but also some opportunities, which were also revealed in an evaluation of students and lectures alike.

Keywords: Teaching R, GUI, Business Application , R Support

Interactive applications written in R to accelerate statistical learning

Chris Wild

Department of Statistics, University of Auckland, New Zealand
https://www.stat.auckland.ac.nz/

Abstract: iNZight and VIT are interactive systems written in R aimed at accelerating statistical learning (in the educational sense). Our premise is that whereas the data world is growing explosively, the rate at which students become exposed to this expanding world is far too slow. We can greatly speed up how quickly students can experience aspects of working with data using software that finesses away choke points and time sinks. Our software caters from beginners to sophisticated multi-variable graphics and modelling (including for complex-survey data). It supports data analysis and resampling-based inference. It underpinned my FurtureLearn MOOC "Data to Insight", gaining an extremely positive reaction from an international, largely adult, audience. It is also used by large numbers of NZ highschool and undergraduate students. iNZight and VIT consist of sets of R packages called by graphical interactive user interfaces provided by Gtk+ via gwidgets and RGtk2. Each major iNZight module corresponds to an R package. Its online version iNZight Lite is a server-delivered application that hooks iNZight's R packages to a Shiny user-interface all wrapped up with R in a docker container. We will talk about acceleration goals and strategies, and challenges in building and maintaining these systems (and seek collaborators).

Keywords: interactive guis, visualisation, dynamic graphics, Shiny, statistics education

Classroom experiments

James Curran

Department of Statistics, University of Auckland, Auckland, New Zealand
http://www.stat.auckland.ac.nz

Abstract: It is often very difficult to motivate the ideas of experimental design and ANOVA to students who have never actually performed an experiment. This is further compounded by having a class size of 250. In this talk I will briefly discuss a simple R based timing experiment that originally arose from research. This experiment was made student-specific, meaning each student would have slightly different results. I will talk about how the students handled this problem and how we used distributed processing to compare their results with the expected result.

Keywords: Experimental Design, Education, Cluster computing

Statistical Methodology 1

Chair: Helle Sørensen

TAM: An R Package for Item Response Modelling

Thomas Kiefer[†], ***Alexander Robitzsch***[‡], ***Margaret Wu***[‡]

[†]*Federal Institute for Educational Research, Innovation and Development of the Austrian School System, Salzburg, Austria*
[‡]*Victoria University, Melbourne, Australia*

https://www.bifie.at/

Abstract: Item response theory (IRT) embeds a set of statistical models that are used while developing instruments for testing and diagnosing subjects, and analyzing the resulting data – for instance, in educational large-scale assessments such as PISA (Programme for International Student Assessment) as well as in clinical and experimental psychology. Items are administered to subjects in order to measure their latent trait (e.g., ability). IRT models can be divided into several model families regarding the distributional assumptions of the latent trait and assumptions about item response probabilities which are (often) motivated by substantive theory.

This talk presents the R package TAM, which provides the necessary tools for the practitioner to apply IRT models. That is, estimates of item characteristics and latent traits are provided amongst other measures like item fit and model fit statistics. There are several (commercial) IRT software packages such as ConQuest, BILOG, IRTPRO or Mplus and several R packages such as mirt, eRm and ltm. Yet, the R package TAM has a considerable processing speed, which is achieved by using an accompanying C++ library. To ease the specification of the broad class of IRT models implemented in TAM, a modelling syntax based on the R package lavaan is introduced.

Keywords: Psychometrics, Item Response Theory, Educational Assessment

gets: General-to-Specific (GETS) Modelling

Genaro Sucarrat

Department of Economics, BI Norwegian Business School, Norway
http://www.bi.no/

Abstract: General-to-Specific (GETS) modelling starts with a General Unrestricted Model (GUM) that is validated against a chosen set of diagnostic tests. Next, multi-path simplification is undertaken by means of backwards elimination, where each regressor-elimination is checked against the chosen diagnostic 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 diagnostic tests or the BaT. Since simplification is undertaken along multiple paths, this usually results in multiple terminal models. An information criterion is then used to choose among them.

The gets package is the successor of AutoSEARCH. The successor is more user-friendly, faster and contains more features. It provides facilities for automated multi-path GETS modelling of the mean and variance of a regression, and Indicator Saturation (IS) methods for the detection and test of structural breaks in the mean. The mean can be specified as an autoregressive model with covariates (an 'AR-X' model), and the variance can be specified as a log-variance model with covariates (a 'log-ARCH-X' model). The four main functions of the package are arx, getsm, getsv and isat.

Keywords: general-to-specific modelling, impulse and step indicator saturation, sir professor david f. hendry oxford

R Package CASA: Component Automatic Selection in Additive models

Thouvenot Vincent

EDF R&D, Clamart, France/ Univ. Orsay, Orsay, France

http://www.math.u-psud.fr/

Abstract: For electricity providers, forecasting electricity demand is a key activity as it is one of the most important entries for production planning and trading on the electricity demand. The development of smart grids and more generally the new data sources coming from new metering infrastructures for energy management, as well as the consumption habits changes, induce a need for new analytics. New statistical methods allowing automatic feature selection and model calibration are thus a necessary tool for many actors of energy markets. We propose here a new automatic methodology based on grouped lasso variable selection.

The R package CASA provides an algorithm which allows efficient selection and estimation of additive model, which achieve a tradeoff between good forecasting performances and low human intervention, by combining Group LASSO and P-Splines estimator, which are efficient in selection and in estimation respectively. Our procedure does not suffer from the traditional bias of a Lasso (and thus grouped Lasso) selector. We develop the main possibilities offer by CASA on the nice case study GEFCom12 public dataset which deals with meteorological covariate selection for local electricity forecasting in US.

Keywords: Electricity forecasting, Group LASSO, Multi-step estimator, P-Spline, Sparse additive model

Dose-response analysis using R revisited

Christian Ritz

Department of Nutrition, Exercise and Sports, University of Copenhagen, Denmark
http://www.nexs.ku.dk

Abstract: Over the last 20 years the open-source environment R has developed into an extremely powerful statistical computing environment. The availability of such a programming infrastructure has in turn fuelled the development of highly sophisticated smaller and larger sub systems for more or less specialized statistical analyses within a number of scientific areas (e.g., the bionconductor suite of packages). One such specialized sub system is provided mainly through the add-on package drc (abbreviation of dose-response curves). Originally, back in 2004, drc was designed as a package for provided model fitting and plotting functionality for very specialized analyses that were routinely carried in weed science.

Over the last decade the package has been modified and extended substantially, mostly in response to numerous inquiries and questions from the user community. Consequently, the package has developed into a very flexible and versatile package for dose-response analysis in general. Applications are found across a wide range disciplines from animal production over medical research to environmental toxicology. Also, the package is used both in academia and industry.

We will briefly outline the key functionality of drc. Furthermore, we will demonstrate some recent major extensions. Finally, we will indicate some directions for future development and research.

Keywords: built-in equations, effective doses, generalized nonlinear regression, log-logistic model, self starter routines

Changepoints over a Range of Penalties using the changepoint package

Kaylea Haynes

Statistics and Operational Research (STOR-i), Lancaster University, United Kingdom
http://www.stor-i.lancs.ac.uk/

Abstract: The changepoint package (Killick and Eckley 2014) is designed to implement various changepoint methods for finding single and multiple changepoints within data. In particular it provides an implementation of the PELT (Pruned Exact Linear Time) algorithm (Killick et al., 2012), which, under certain conditions, has a computational cost linear in the number of data points. The main challenge with this method is it requires a penalty value to be chosen, and this choice can substantially affect the accuracy of the estimated changepoints. To overcome this problem we have developed a new method, CROPS (Changepoints over a Range of PenalteS), (Haynes et al., 2014), which allows one to obtain optimal changepoint segmentations of data sequences for all penalty values across a continuous range. The computational complexity of this approach can be linear in the number of data points and linear in the optimal segmentations for the smallest and largest penalty values. This algorithm is implemented in the latest release of the changepoint package. This talk will introduce our new algorithm CROPS which we will illustrate on both simulated and empirical data sets using the changepoint package. We will also highlight some of the new features of the recently updated changepoint package.

Keywords: PELT, Segmentation , Dynamic Programming

Machine Learning 1

Chair: Poul Svante Eriksen

Rapid detection of spatiotemporal clusters

Markus Loecher

Dept. of Economics, Berlin School of Economics and Law, Germany
http://www.hwr-berlin.de/en/

Abstract: Identifying spatially and temporally defined hotspots of activity is an important tool for many technologies and scientific disciplines. This task can often be reduced to detecting over-densities in space relative to a background density. This relative density estimation is framed as a binary classification problem. An integrated hotspot visualizer is presented which allows the efficient identification and visualization of clusters in one environment. Many real data exhibit multiple clusters at many different scales at various angles. Rather than aggregating to a given spatial partition, algorithms such as classification trees can find regions with high and low densities or rates w.r.t. some baseline. We succeed in identifying rectangular "leafs" of high or low incidence of one of the two classes by growing trees on various rotations of the data. While a binary recursive tree partitions space exhaustively and in that sense is not naturally suited as a hotspot detector, we depart from this traditional view and simply retain the most "interesting boxes", typically those with an average rate above a chosen threshold.

Keywords: classification trees, scan statistic, density estimation, hotspots

Scalable distributed random-forest in R

Arash Fard, Vishrut Gupta

HP DistributedR

https://github.com/vertica/DistributedR/

Abstract: Random Forest and Gradient Boosting Machine are widely used methods for building ensembles of decision trees for classification and regression. However, the training phase of these algorithms can easily take from hours to days to complete on even small datasets. Due to their complex nature, few, if any, distributed implementations of decision tree construction algorithms are available in open source.

In this talk, we will explain two different versions of distributed Random Forest that we have implemented using the open source "distributedR" package. In the first implementation, each server builds independent trees and then aggregates them in a final model. While this approach shows good performance, it is limited by the fact that the full data has to be replicated on each machine. We will present a second implementation, where each individual tree is created in parallel across all machines. Our experience shows that this new distributed Random Forest package scales linearly and can easily process multi-gigabyte data-sets in R. More importantly, the same technique can be used to implement distributed version of other tree ensemble methods such as Gradient Boosting Machine.

Keywords: distributed computing, machine learning, big data, random forest

Multivariate analysis of mixed data: The PCAmixdata R package

Marie Chavent, *V. Kuentz, A. Labenne and J. Saracco*

Inria Bordeaux Sud-Ouest and University of Bordeaux, France

http://www.inria.fr/en/teams/cqfd

Abstract: Mixed data type arise when observations are described by a mixture of numerical and categorical variables. The R package PCAmixdata extends standard multivariate analysis methods to incorporate this type of data. The key techniques included in the package are PCAmix (PCA of a mixture of numerical and categorical variables), PCArot (rotation in PCAmix) and MFAmix (multiple factor analysis with mixed data within a dataset). A synthetic presentation of the three algorithms will be provided and the three main procedures will be illustrated on real data composed of four datasets caracterizing conditions of life of cities of Gironde, a south-west region of France.

Keywords: Multivariate data analysis, Mixture of numerical and categorival variables, Multi-group data, Principal Component Analysis, Rotation

PPforest

Natalia da Silva

Department of Statistics, Iowa State University, Iowa, United States of America
http://www.stat.iastate.edu

Abstract: We present an R package called PPforest available on CRAN. This package implements a projection pursuit classification random forest. A random forest is an ensemble learning method, built on bagged trees. The bagging provides power for classification because it yields information about variable importance, predictive error and proximity of observations. This research adapts the random forest to utilize combinations of variables in the tree construction, which we call the projection pursuit classification random forest (PPforest). In a random forest each split is based on a single variable, chosen from a subset of predictors. In the PPforest, each split is based on a linear combination of randomly chosen variables. The linear combination is computed by optimizing a projection pursuit index, to get a projection of the variables that best separates the classes. The PPforest uses the PPtree algorithm, which fits a single tree to the data. Utilizing linear combinations of variables to separate classes takes the correlation between variables into account, and can outperform the basic forest when separations between groups occur on combinations of variables. Two projection pursuit indexes, LDA and PDA, are used for PPforest.

Keywords: Random forest, projection pursuit, supervised classification, exploratory data analysis, data mining

Visualisation 1

Chair: Di Cook

Reordering and selecting continuous variables for scatterplot matrices

Katrin Grimm

Department of Computeroriented Statistics and Data Analysis, University of Augsburg, Germany

http://www.rosuda.org

Abstract: The identification and visualisation of correlation patterns can be challenging for large, new datasets. Standard visualisation techniques like scatterplot matrices become less informative as the number of variables increases. To get an initial overview two questions can be considered:

1. What is the general correlation structure?
2. Which q of all p continuous variables are especially highly correlated and how do the bivariate dependencies look?

The first question can be answered visually by corrgrams. By using a new ordering of the variables, based on the angles of the first two eigenvectors of the correlation matrix, correlation patterns amongst the variables can be emphasised. Reordering can also be helpful for answering the second question.

The talk presents a further approach, selecting groups of variables which are all highly correlated with one another. Only variable combinations are considered, which are adjacent in the reordered list of variables. This avoids having to check all possible combinations to reduce the complexity of the calculation. The talk includes a formal description of the approach and a practical example to show what results are obtained in practice.

Keywords: Correlation, Variable selection, Variable reordering , Visualisation , Scatterplot matrices

R-package to assess and visualize the calibration of multiclass risk predictions

Kirsten Van Hoorde

Open Analytics NV

http://www.openanalytics.eu

Abstract: Risk prediction models for diagnostic or prognostic outcomes are useful tools for clinical decision support, personalized healthcare and shared decision making. Clinical problems are often dichotomized for analysis - e.g. a benign or malignant ovarian tumor – although the underlying problem is multiclass or multinomial – e.g. a benign, borderline or invasive ovarian tumor. A more detailed differentiation can help to optimize (patient) management as well as patient survival.

Calibration is an important aspect when evaluating risk prediction models, i.e. whether the predicted risks correspond to the observed probabilities. The optimal use of risk prediction models relies on reliable risk estimation.

For binary outcomes, several tools exist to assess different aspects of model calibration (e.g. calibration-in-the-large and calibration plots). We extended these tools towards multinomial risk prediction models [1]. We proposed multinomial calibration plots which give a visual summary of the calibration performance [1]. Furthermore, we presented generic tools to assess the calibration of multiclass risk models [2].

The integration of these different multiclass calibration tools into a new R package multiCalibration (multinomial/multiclass calibration) fills the gap of multinomial calibration tools and allows the user to assess and visualize calibration of multiclass risk predictions irrespective of the used modeling technique.

Keywords: multiclass, calibration, risk prediction, calibration plot, risk model

tmap: creating thematic maps in a flexible way

Martijn Tennekes

Statistics Netherlands, Heerlen, The Netherlands

http://www.cbs.nl

Abstract: A thematic map is a geographical map in which statistical data are visualized. The theme refers to the statistical phenomena that is shown, such as the unemployment rate at municipal level. The best known thematic map type is the choropleth, where regions are coloured according to a statistical variable, for instance unemployment rate or population density. Another popular thematic map type is the bubble map, in which the sizes of the bubbles are defined by a statistical variable, for instance city population.

With the tmap package, thematic maps can be generated with great flexibility. The syntax for creating plots is similar to that of ggplot2. A thematic map is created by stacking layers, for instance one for colouring municipalities, one for thick borders of higher level regions, and one for text labels. It is also possible to create small multiples. Layout settings such as legend positioning and margins can be specified for particular shape objects by layout themes.

The standard work flow that is needed to create a thematic map is embedded in tmap by several convenient functions, e.g., for reading ESRI shape files, setting the proper projection, appending data, and calculating densities from absolute values.

Keywords: thematic maps, data visualization, official statistics

The dendextend R package for manipulation, visualization and comparison of dendograms

Tal Galili

Tel Aviv University

http://www.math.tau.ac.il/index.php?Itemid=27

Abstract: A dendrogram is a tree diagram which is often used to visualize a hierarchical clustering of items. Dendrograms are used in many disciplines, ranging from Phylogenetic Trees in computational biology to Lexomic Trees in text analysis. Hierarchical clustering in R is commonly performed using the hclust function. When a more sophisticated visualization is desired, the hclust object is often coerced into a dendrogram object, which in turn is modified and plotted. The dendextend R package extends the palette of base R functions for the dendrogram class, offering easier manipulation of a dendrogram's shape, color and content through functions such as rotate, prune, color_labels, color_branches, cutree, and more. These can be plotted in based R and ggplot2. dendextend also provides the tools for comparing the similarity of two dendrograms to one another: either graphically (using a tanglegram plot, or Bk plots), or statistically (with Cophenetic correlation, Baker's Gamma, etc) - while enabling bootstrap and permutation tests for comparing the trees. The dendextendRcpp package provides C++ faster implementations for some of the more computationally intensive functions.

Keywords: dendrogram, visualization, clustering, hierarchical clustering, ggplot2

Kaleidoscope 6

Chair: Susan Holmes

The METACRAN experiment

Gabor Csardi

Department of Statistics, Harvard University, Cambridge, MA, USA

http://statistics.fas.harvard.edu

Abstract: CRAN is serving us well as the main point of distributing R packages, and has been one of the main reasong behind the success of R.

METACRAN is an experiment to provide additional services on top of the CRAN infrastructure:

- A searchable code mirror, with diffs between package versions, and the ability to create personalized versions of R packages.
- A database and API of CRAN metadata.
- A package search engine and web site for package discovery.
- A package manager.
- A continuous integration wrapper to build and check R packages with various R versions.
- A database and API of CRAN package downloads from the cloud CRAN mirror.

In this presentation I briefly introduce these services.

Note that METACRAN is not a CRAN project, and it is developed and maintained independently of CRAN.

Keywords: CRAN, R packages, discussion

Using R in photobiology

Pedro J. Aphalo

Department of Biosciences, University of Helsinki, Finland
http://www.helsinki.fi/biosciences/

Abstract: The packages in the R4Photobiology suite implement data import, calculations and plotting related to handling of spectral data as used in photobiology, plus some additional functions for day length and sun position calculations. The main aim is to make it easier for biologists to quantify and describe the visible and ultraviolet radiation conditions used in experiments or monitored in nature, in a standardized and consistent way. In the spirit of reproducible research the interface, is not visual or menu based, and all functionality is available in knitr scripts. My design is based on the idea of mirroring as much as possible the "concepts" used in the user domain, photobiology. The design is based on a core package and additional packages providing data examples and functions for specific fields, in the hope that contributions will in the future expand the usefulness of the suite. Plotting functions are built using ggplot2 as a basis, but they are grouped in a separate package, making parallel future implementations based on gvis or other plotting "systems" possible. An object oriented design is used. The capabilities of the packages will be described, and the design of the interface and its implementation, will be presented.

Keywords: New packages, Photobiology, Spectral data, User interface design, Implementation in R

Industrial Big Data Analytics for Wind Turbines

Sven Jesper Knudsen, *Martin Qvist, Kim-Emil Andersen*

Vestas Wind Systems A/S, Vestas Performance and Diagnostics Centre

Abstract: Vestas is forerunner in using industrial big data and R plays a key role. We will show how R is used as a compute engine in the scale-out processing capabilities of Hive, and share lessons learned from a technical, a user, and a business perspective.

Vestas monitor more than 25.000 wind turbines. All of these are fitted with a multitude of sensors, and they spin off a vast amount of data. For a decade, engineers have in turn created analytics to help predict, say, when a crucial part require repair. This effort has significantly reduced unscheduled downtime and lost wind energy production.

Taking this to the current era, we have utilized Hive on Vestas' supercomputer, FireStorm. This provides a query language not unlike SQL that allows us to combine data sets that we have never been able to do before. Data scientists can infuse R mappers and/or reducers and hence combine the scale-out processing capabilities with the analytic capabilities of R. It is still a bit tech-savvy, and sometimes hard to control in lack of a query optimiser, though.

Keywords: Industrial Big Data, Monitoring analytics, Performance analytics, Meso scale climate data , Hive

The Network Structure of R Packages

Andrie de Vries

Revolution Analytics

http://www.revolutionanalytics.com/

Abstract: Over the past decade, R evolved from being a somewhat specialized platform supporting statisticians to a widely used tool at the center of the new developments in data science.

Much of the growth in the capabilities of R is due to the success of R's package repository system. CRAN (> 6,000 packages) and BioConductor (> 900 packages) are the two primary repositories. These repositories are managed independently with different release cycles and different conformance policies.

In this presentation, I use statistical network theory and the algorithms implemented in various R packages including igraph and miniCRAN to analyze and visualize the connectivity structure of packages in CRAN and BioConductor. I also test the hypothesis that the different management policies of the two repositories are reflected in the properties of their graph networks.

Keywords: graph networks, community detection, pagerank, igraph, miniCRAN

Teaching 2

Chair: Helle Sørensen

Web Application Teaching Tools for Statistics Using Shiny and R

Gail Potter, *Jimmy Doi, Peter Chi, Jimmy Wong and Irvin Alcaraz*

Department of Statistics, California Polytechnic State University, San Luis Obispo, CA, U.S.A.

http://statistics.calpoly.edu/

Abstract: Technology plays a critical role in supporting statistics education, and student comprehension is improved when simulations accompanied by dynamic visualizations are employed in the class. Several web-based applets programmed in Javascript are publicly available (e.g. www.rossmanchance.com). These provide a user-friendly interface which is accessible and appealing to students in introductory statistics courses. However, many statistics educators are not fluent in Javascript so are unable to tailor these apps or develop their own. We provide an introduction to Shiny, a web application framework for R created by RStudio, which facilitates applet development for educators who are familiar with R. We illustrate the utility, convenience, and versatility of Shiny through twenty freely available apps covering a range of topics and levels (found at http://statistics.calpoly.edu/shiny). For example, one app compares the robustness of the Mann Whitney U-Test to the t-test, and another graphically demonstrates the Probability Integral Tranform and the Accept-Reject Algorithm for random variable generation. Our source code is publicly available so that educators may tailor our apps as desired. As the movement towards online education and web-based teaching tools gains momentum, educators who adapt to cutting-edge technology such as Shiny will remain at the forefront.

Keywords: Shiny, web application, statistics education, teaching tools, applets

Teaching R in (an online) class

Jonathan Cornelissen

DataCamp

http://www.datacamp.com

Abstract: Over 120,000 people have started a free R course on DataCamp so far (over 50k the "introduction to R": https://www.datacamp.com/courses/introduction-to-r). We have received numerous requests from Professors who use our free introduction to R course in their classes to get more insight in student's performance. Therefore, we will launch a new interface that gives Professors insight in how their students are learning R. Additionally, we'll give a brief intro on how to create courses on DataCamp for both the traditional interface and the new swirl interface https://www.datacamp.com/swirl-r-tutorial. Finally, we will share some key insights based on analyzing the data of over 100k students learning R.

Keywords: Teaching R, Online learning, learning analytics, DataCamp, swirl

Teaching R using the github ecosystem

Colin Rundel

Department of Statistical Science, Duke University, USA

https://stat.duke.edu/

Abstract: In this talk we will discuss an R based Master's level statistical computing course taught for the first time at Duke University in the Fall of 2014. The talk will focus on ways in which statisticians can adopt the tools and best practices of software engineers and other working practictioners to teach and improve the computing abilities of our students. We will focus in particular on the adoption of github and connected services like Travis-CI as a platform for teaching students how to produce R based analyses that are of high quality, reproducible, and throughly tested while working in a team-based collaborative environment. We will discuss in detail course structure and logistics as well as give examples from the case studies used in the course.

Keywords: Teaching, Reproducibility, Testing, github, Continuous integration

Using R, RStudio, and Docker for introductory statistics teaching

Mine Cetinkaya-Rundel

Department of Statistical Science, Duke University, USA

https://stat.duke.edu/

Abstract: Docker is a relatively new Linux container technology that can be used as a lightweight form of virtualization. At Duke University we have been using Docker containers to provide students with anywhere access to RStudio Server via the campus-wide authentication system. This solution eliminates the need for students to locally install R and RStudio, and hence makes getting started with R considerably easier for students who are brand new to statistics and computing. This solution also provides secure self-contained virtualized environments for each student with specified quotas for system resources. Additionally, it simplifies the maintenance and support while allowing for deployment across multiple heterogeneous physical systems. In this talk we will discuss implementation and benefits of using Docker and RStudio Server as an environment for teaching R at the introductory level.

Keywords: Docker, RStudio, teaching

Statistical Methodology 2

Chair: James Curran

seasonal: An X-13 interface for seasonal adjustment

Christoph Sax

University of Basel

https://wwz.unibas.ch/personen/profil/person/sachs/

Abstract: 'seasonal' is an easy-to-use and full-featured R-interface to X-13, the newest seasonal adjustment software developed by the United States Census Bureau. X-13 is a free command line software, written in Fortran, which is used by many statistical agencies. It offers a large toolkit to seasonally adjust time series, including fully automated methods.

With 'seasonal', it is possible to access the almost complete syntax of X-13, using a simple R command. It also offers a simple and flexible mechanism to read almost all output from X-13. Beside interfacing to X-13, 'seasonal' includes the capability to adjust for user defined holidays, such as Chinese New Year or Indian Diwali. 'seasonal' also includes a 'shiny'-based graphical user interface, which can also be explored online: www.seasonal.website.

Keywords: seasonal adjustment, time series, official statistics, interface, shiny

Estimating the Linfoot correlation in R

Sören Möller

Epidemiology, Biostatistics and Biodemography, Department of Public Health, University of Southern Denmark, Odense, Denmark

http://www.sdu.dk/en/Om_SDU/Institutter_centre/Ist_sundhedstjenesteforsk/Forskning/Forskningsenheder/Epidemiologi/

Abstract: In 1957 E. H. Linfoot introduced his measure of correlation based on the mutual information. It is one of the few dependence measures known to fulfil all seven of Renyi's (1959) desirable properties for a dependency measure. Still, the Linfoot correlation is not widely used in practice, as it is relatively hard to estimate.

We demonstrate how the Linfoot correlation can be estimated efficiently in R. We present various non-parametric approaches using numerical integration of kernel densities and k-nearest neighbour estimates. Moreover, we investigate a parametric approach fitting Archimedean copulas to the data. We implement these methods for R in our forthcoming Linfoot package.

We apply these methods to simulated data with known dependency as well as real world data, and compare the Linfoot correlation to the classical Pearson, Spearman and Kendall correlations to investigate the relationship between these measures of dependency. Furthermore, we compare our different approaches of estimation with each other, to check the robustness of our estimates and their dependency on the choice of copula.

Keywords: dependence measure, Linfoot correlation, mutual information, copula

Seasonal Adjustment with the R packages x12 and x12GUI

Alexander Kowarik

Methods, Statistics Austria, Vienna, Austria

http://www.statistik.gv.at

Abstract: The X-12-ARIMA Seasonal Adjustment program of the U.S. Census Bureau extracts the different components of a monthly or quarterly time series. It is the state-of-the-art technology for seasonal adjustment used in many statistical offices. The procedure makes additive or multiplicative adjustments and creates an output data set containing the adjusted time series and intermediate calculations. The original output from X-12-ARIMA is somehow static and it is not always an easy task for users to extract the required information for further processing. The R package x12 provides wrapper functions and an abstraction layer for batch processing X-12-ARIMA. It allows summarizing, modifying and storing the output from X-12-ARIMA within a well-defined class-oriented implementation. On top of the class-oriented implementation the graphical user interface allows access to the R package x12 without demanding too much R knowledge. Users can interactively select additive outliers, level shifts and temporary changes and see the impact immediately. Having the powerful X-12-ARIMA Seasonal Adjustment program available directly from within R, as well as containing the new facilities for marking outliers, batch processing and change tracking, makes the package a potent and functional tool.

Keywords: seasonal adjustment, time series, outlier detection

frailtyHL: R package for variable selection in general frailty models for various survival data

Il Do Ha[†], ***Maengseok Noh***[†], ***Donghwan Lee***[‡], ***Youngjo Lee***[§]

[†]*Department of Statistics, Pukyong National University, Busan, South Korea,*
[‡]*Department of Statistics, Ewha Womans University, Seoul, South Korea*
[§]*Department of Statistics, Seoul National University, Seoul, South Korea*

http://www.pknu.ac.kr

Abstract: Variable selection methods using a penalized likelihood have been widely studied in various statistical models. However, in semi-parametric frailty models, these methods have been relatively less studied because the marginal likelihood function involves analytically intractable integrals, particularly when modeling mullti-component or correlated frailties. The frailtyHL R package (Ha, Noh and Lee, 2012) can be used for fitting semi-parametric frailty models using h-likelihood (Lee and Nelder, 1996), which does not require intensive numerical methods to find the marginal likelihood. In this talk we introduce an updated frailtyHL package via a penalized h-likelihood for variable selection of fixed effects in a general class of semiparametric frailty models, in which random effects may be shared, nested, or correlated. Here we allow three penalty functions (least absolute shrinkage and selection operator [LASSO], smoothly clipped absolute deviation [SCAD], and HL) in our variable selection procedure. We illustrate the use of our package with the well-known practical data sets, and compare our results with alternative R-procedures.

Keywords: Frailty models , Penalized h-likelihood , LASSO, SCAD, Variable selection

Machine Learning 2

Chair: Poul Svante Eriksen

Massive Online Data Stream Mining using R and MOA

Jan Wijffels

BNOSAC, Brussels, Belgium

http://www.bnosac.be

Abstract: MOA (http://moa.cms.waikato.ac.nz) is the most popular open source framework for data stream mining.

With our new package RMOA (http://cran.r-project.org/web/packages/RMOA/index.html), which focusses on building streaming classification and regression machine learning models, R users can now easily interface their data to MOA in order to build, evaluate and score classification models like hoeffding trees, naive bayes rules, knn models, ensemble models like bagging, boosting, stacking on streaming data while limiting RAM usage. Next to streaming classifications, RMOA sets up an R friendly interface to the regression facilities of MOA such that evaluation of streaming perceptrons and the use of stochastic gradient descent iterative regression modelling on streaming data becomes a breeze.

The usage of RMOA will be showcased on a Turtlebot Robot (http://turtlebot.com) by using the rosR bridge between R and ROS (http://journal.r-project.org/archive/2013-2/dietrich-zug-kaiser.pdf).

Keywords: Massive Online Analysis, Streaming modelling, RAM, Machine Learning, Robotics

forestFloor: a package to visualize and comprehend the full curvature of random forests

Søren Havelund Welling

DTU, Compute; NovoNordisk, Insulin Pharmacology Research

http://www.compute.dtu.dk/

Abstract: forestFloor is an add-on to the randomForest[1] package. It enables users to explore the curvature of a random forest model-fit. In general, for any problem where a random forest have a superior prediction performance, it is of great interest to learn its model mapping. Even within statistical fields where random forest is far from standard practice, such insight from a data driven analysis can give inspiration to how a given model driven analysis could be improved.

forestFloor is machine learning to learn from the machine!

A mapping-function of a random forest model is most often high dimensional and therefore difficult to visualize and interpret. However, with a new concept, feature contributions[2-3], it is possible to split the random forest mapping-function into additive components and understand the full curvatures. Hereby the forestFloor package provides a great extended functionality, compared to the original partial dependence plot provided in the randomForest package. To explore the curvature of random forests through series of 2D/3D plots with use of color gradients is fun and quite intuitive. forestFloor relies amongst others on Rcpp, rgl and kknn packages to produce visualizations fast and smoothly.

Keywords: random forest, machine learning, QSAR, visualizations, exploratory analysis

Machine Learning for Internal Product Measurement

Douglas Mason

Twitter, Inc.

http://www.twitter.com

Abstract: Most people think of machine learning applications in user-facing features such as recommendation engines. However, machine learning techniques are also tantamount for rigorous internal product measurement. In this talk we will compare different common approaches such as logistic regressions, random forests, and deep learning neural nets for product measurement, their pros and cons, and how we can use them to inform research-based product design.

Keywords: Machine Learning, Logistic Regression, Product Measurement

h2oEnsemble for Scalable Ensemble Learning in R

Erin LeDell

H2O

http://h2o.ai/

Abstract: Ensemble machine learning methods are often used when the true prediction function is not easily approximated by a single algorithm. There is an implicit computational cost to using ensemble methods, since they require the training of multiple base learning algorithms. Practitioners may prefer ensemble algorithms when model performance is valued above other factors such as model complexity and training time. We present the `h2oEnsemble` R package, which reduces the computational burden of ensemble learning while retaining superior model performance. This R interface provides easy access to scalable ensemble learning.

The H2O Ensemble software implements the Super Learner, or stacking, ensemble algorithm, using distributed base learning algorithms from the open source machine learning platform, H2O. The following base learner algorithms are currently supported in `h2oEnsemble`: Generalized linear models with elastic net regularization, Gradient Boosting (GBM) with regression and classification trees, Random Forest and Deep Learning (multi-layer feed-forward neural networks).

Keywords: machine learning, ensemble learning, parallel computing, distributed computing, cross-validation

Visualisation 2

Chair: Hadley Wickham

Plotting data as music videos in R

Thomas Levine

csv soundsystem

http://tools.ietf.org/html/rfc4180

Abstract: I have been experimenting with plotting data in the form of music videos with the goal of plotting complex, multidimensional data in intuitive ways. I will conceptually discuss how I map data to video elements and musical elements, and I will give practical demonstrations of synthesizing music and video in R. Aside from learning how to make music videos in R, attendees can expect to learn about some of the more esoteric aspects of base R plots.

Keywords: plot, visualization, sonification, base-graphics

NaviCell Web Service for Network-based Data Visualization

Eric Bonnet

Institut Curie, Paris, France

http://sysbio.curie.fr/

Abstract: NaviCell Web Service (https://navicell.curie.fr/pages/nav_web_service.html) is a tool for biological networks-based visualization of "omics" data. It implements several data visual representation methods, including the novel map staining technique for grasping large-scale trends in numerical values (such as whole transcriptome) projected on top of a pathway map. NaviCell Web Service allows combining visualization of different data types. It is scalable, based on Google Maps technology allowing working with large pathway maps, containing thousands of nodes and using semantic zooming principles. NaviCell Web Service can be applied to pathway maps of different types represented in various formats in different application fields. The web service provides a server mode, which allows automating visualization tasks and retrieve data from maps via RESTfull (standard HTTP) calls. Bindings to different programming languages are provided (Python, R, Java). We illustrate the purpose of the tool with several case studies using different pathway maps, with an emphasis on the R package R-NaviCell (https://github.com/eb00/R-NaviCell).

Keywords: biological networks, visualization, high-throughput data

Easy visualizations of high-dimensional genomic data

Laure Cougnaud

Open Analytics, Antwerpen, Belgium
http://www.openanalytics.eu/

Abstract: The high dimensionality of genomic data makes the task of extracting and representing the main patterns and features of the data quite complex. This is especially the case for an experiment with many covariates or biological conditions (treatments).

The talk will focus on known and less well-known visualizations of such genomic data, especially in the context of the analysis of microarray experiments.

Representations of spectral maps, techniques based on neighbouring methods such as t-SNE (T-Distributed Stochastic Neighbor Embedding) or traditional linear discriminant analysis will be adressed in the genomic context. A dedicated R package will be presented which uses the ExpressionSet class (for easier storage of genomic data and its annotation) to integrate well with the Bioconductor workflow. The visualization logic is based on proper aesthetic mappings with ggplot2 whereas interactivity is offered via RMarkdown and Shiny.

Keywords: genomics, unsupervised analysis, high-dimensional, visualization

The gridGraphics Package

Paul Murrell

Department of Statistics, The University of Auckland, Auckland, New Zealand
https://www.stat.auckland.ac.nz/

Abstract: Many important R packages provide graphics functions based on the "base graphics" system (they depend on the 'graphics' package), but several other important R packages, including 'lattice' and 'ggplot2', provide functions based on the alternative "grid graphics" system (they depend on the 'grid' package). Unfortunately, the base graphics system and the grid graphics system do not integrate well with each other and this can lead to a number of problems: users can become confused if they accidentally combine functions from the two graphics systems; plots produced with base graphics do not have access to the more powerful customisation features of the grid graphics system; and there can be duplication of effort when developers produce two versions of the same plot – one based on 'graphics' and one based on 'grid'.

This talk describes the new 'gridGraphics' package, which provides functions to convert any plot that has been drawn with functions based on the 'graphics' package to an identical plot drawn with the 'grid' package. The talk will explain the background to the problem of having two distinct graphics systems, show how to use the 'gridGraphics' package, and demonstrate the value of the new package via a short case study.

Keywords: plots, graphics (package), grid (package), gridGraphics (package)

Part III

Lightning Talks

Lightning talks

Chair: James Curran

An implementation of the SAEM algorithm for left-censored data

Raphaël Coudret

Open Analytics NV, Antwerp, Belgium

http://www.openanalytics.eu/

Abstract: When studying a model with unknown real variables, the maximum likelihood estimator is very popular. For some models, it is not possible to have the expression of the likelihood. To tackle this issue, the EM algorithm was introduced. This iterative algorithm relies on the expectation of another likelihood, given some hidden random variables, but this expectation can also be impossible to find.

The SAEM algorithm is an improvement of the EM algorithm. It estimates the desired expectation and produces a sequence of estimates of the unknown real variables. If we could compute the likelihood function for each element of this sequence, it would converge toward a local maximum of this likelihood function. See Delyon, Lavielle and Moulines (1999, Theorem 7).

We provide an R package containing an implementation of the SAEM algorithm for models with left-censored observations. This kind of models is commonly considered in pharmaceutical companies to handle inaccurate measures when the studied quantities are too small.

The SAEM algorithm also works for some models with a known likelihood function. That is why our package includes dynamic outputs to observe the convergence of interest. They can be used to check the implementation and to choose optimal parameters.

Keywords: Maximum likelihood estimator, Stochastic approximation, Censoring, Expectation-maximization algorithm

Crowdsourced Data Processing with MTurkR

Thomas J. Leeper

Department of Political Science, Aarhus University, Aarhus, Denmark
http://ps.au.dk/

Abstract: This talk will introduce the use of the Amazon Mechanical Turk (MTurk) crowdsourcing platform as a resource for R users, focusing on ways that MTurk can help deal with messy data problems. The talk will introduce the MTurkR package <http://cran.r-project.org/web/packages/MTurkR/index.html>, which connects useRs to the human intelligence provided by MTurk crowd workers who can code data, images, and text, transcribe audio, video, or other machine-unreadable data, and thereby expand the capabilities of R beyond machine computation alone. The talk will focus on concrete applications and point the audience to further resources for creating and managing complex crowdsourcing tasks through R.

Keywords: crowdsourcing, processing, coding, transcription

Development and validation of statistical models for occupancy detection in an office building

Luis Miguel Candanedo Ibarra

Faculté Polytechnique, Service de Thermique et Combustion, Université de Mons, Belgium

https://portail.umons.ac.be/fr/universite/facultes/fpms/recherche/gr_ser/serv_term_comb/pages/default.aspx

Abstract: Real time building occupancy detection can be used to minimize the energy consumption in buildings by controlling the heating, cooling and lighting systems. In this work, experimental data from temperature, humidity, CO2 and light sensors have been used together with a digital camera to determine the truth occupancy status of an office room. The recorded labeled data is fed to supervised statistical learning models for classification. For this task, the open source tool R was used for data processing and model training. The results show very encouraging results. High accuracies were found in the data testing sets using random forest models and linear discriminant analysis. Also different feature combinations are compared. This is interesting since different combination of sensors exits in buildings. This work is part of the NEED4B consortium activities and has been co-financed by the European Commission through the Seventh Framework program.

Keywords: Occupancy detection, SVMs, random forests, classification

Drat: Package Repositories Made Easy

Dirk Eddelbuettel

Debian and R Projects

Abstract: The R package ecosystem is one of the cornerstones of the success seen by R. Support for multiple repositories is built into R, but (beyond BioConductor) not widely used.

The drat package allows easy creation of repositories for package authors, and equally easy deployment of repositories by package users. In particular, GitHub can be used as its "built-in" optional web presence for a (git) repository is ideally suited for hosting an R package repository (requiring only a webserver).

We illustrate several test cases: newer packages under development, packages not matching other repository requirements and of course local repositories accessibly only with a research group or department.

Keywords: Packages, Programming, Distribution, Infrastructure

Interrogating Six Large Gene Expression Datasets of Normal Human Brains with RUVcorr/R-Shiny

Saskia Freytag

Walter+Eliza Hall Institute

http://www.wehi.edu.au/

Abstract: Publicly available resources on gene expression in normal human brains can be used to identify gene networks that pertain to diseases of interest and can identify genes that are likely to be involved in the causal mechanism. However, clinicians and biologists are often intimidated by the bioinformatics and R-programming knowledge required to analyze these large datasets. For these reasons we created an interactive R-shiny tool that allows interrogation of six large gene expression studies of the developed and developing human brain. We ensured that datasets were comparable by standardizing gene nomenclature. Furthermore, the user's research interest informs a data-driven cleaning procedure, removal of unwanted variation, shown to lead to greater consistency across different studies. Our tool currently uses the cleaned datasets to visualize networks of genes that the user is interested in and compares these across studies. Additionally, it can be used to rank candidate genes according to their likelihood of involvement in the disease of interest. Future work will expand this tool to include more sophisticated analysis methods such as spatio-temporal network analyses as well as incorporating other omics datasets on the human brain.

Keywords: R-Shiny, Gene Expression, Brain, Data Cleaning, Visualization

Introducing R as first programming

Soma Datta

Northeastern State University Department of mathematics and Computer Science
http://academics.nsuok.edu/mathematics/MathCSHome.aspx

Abstract: The use of computers has become an integral part of everyone's life. Some high schools have introduced programming as a part of their curriculum. Programming languages that are taught in high schools are typically C++, Java, etc. These languages have their own advantages but have relatively intricate syntax, causing resistance toward programming. This paper proposes that an introduction to 'R' from the middle school level would enhance the development of analytical and logical thinking. Students would also develop an interest in Science, Technology, Engineering, & Mathematics (STEM). This paper presents how R can help middle and high school students' analytical development, logical thinking, and develop a love for coding. The hypothesis of this pedagogy is programming can be for all ages and be learnt by themselves with minimal given tools. To prove the hypothesis, this approach will be implemented by introducing R from middle school to college freshmen, with no prior programming knowledge. This study will be evaluated by teaching R and Visual Basic to different groups of students. The methodology to be used will be a modification of the Self Organization Learning Environment (SOLE) concept.

Keywords: First program, R for beginners, Easy to code, Learn R, STEM

Lotka's Law Package

Alon Friedman

School of information, University of south Florida, Tampa, USA
http://www.usf.edu

Abstract: Research suggests that in order to identify a scientific domain, we need to follow the activities of that domain's researchers. The field of informatics addresses the science of processing data for storage using a strong relationship with mathematics to examine any domain. We set out to find whether R can support the examination of common models of informatics by measuring scholarly activities in a particular domain. Researchers have found that the most common laws relied upon by informatics researchers are Bradford's Law, Zipf's Law, and Lotka's Law. Bradford's Law examines journal productivity, Zipf's law measures word frequency and Lotka's Law investigates author productivity. Researchers have implemented Bradford's Law and Zipf's law using R to advance their studies. However, we found that Lotka's Law has not been examined or used under R. In order to address this gap, we developed a Lotka's Law package in R. We found that the advantage of using R in analyzing Lotka's Law is the ability to reexamine Pao's testing procedure. We will present our development of a Lotka's Law package in R and demonstrate how to visualize the results.

Keywords: Informatics, Domain analysis, Lotka's Law, Author productivity, Lotka Law Package

Precipitation extreme value statistics application

Berry Boessenkool

Institute of Earth and Environmental Science, Potsdam University, Germany

http://www.geo.uni-potsdam.de/geoecology.html

Abstract: Precipitation intensity quantiles (P) usually rise exponentially with air temperature (T). This follows the Clausius-Clapeyron relationship for potential air moisture content. Above a certain temperature however, the P-T-relationship is reversed. This drop could be dictated by meteorological properties like limited moisture availability. It could also originate from the fact that the number of observations at high temperatures is small. Possible extreme rainfall intensities may simply not yet be recorded, as high quantiles rise with sample size towards an asymptotic convergence with the theoretical value.

Using R and the package lmomco, we fitted 17 distributions to the precipitation intensity dataset. With the distribution functions closest to the observed data defined as "real distributions" and synthetic random samples of different sizes, simulations confirm our hypothesis that the quantile drop is a statistical artefact. If appropriate distributions are fitted to the samples, their quantiles serve as a non-biased estimator even in small samples. These parametric quantiles show an unabated increase in precipitation intensity with temperature, which is important for flood risk calculation.

Within our research, I created an R package on github, wrote my Master's thesis and am about to present at the EGU general assembly.

Keywords: precipitation intensity, temperature, quantiles, simulations, extreme value stats

Predicting the NCAA Basketball Tournament for Fun and Profit

Jonathan Arfa

Magnetic

http://www.magnetic.com/

Abstract: We will walk through our experience using R to build datasets and models as part of Kaggle's "March Machine Learning Mania 2015" competition, where the task was to accurately predict the outcomes of games in the 2015 NCAA (a university-level athletic association in the USA) basketball tournament. We will go into detail regarding data manipulation, modeling decisions (specifically how the competition's scoring rules informed our choice of model), pitfalls we encountered with bad data and results that were "too good to be true", and what changes we would make for next year's competition.

Our model ended up in the top 25% of competitors after all 63 games were played. With so few games there is of course a lot of randomness in the final ranking, but as people who didn't know much about basketball before this tournament we consider this to be a very good outcome.

Keywords: sports, Kaggle, machine learning, dplyr, caret

R: fast and big strategies.

Adolfo Alvarez

Analyx, Poznan, Poland.

http://www.analyx.com

Abstract: This lightning talk will provide a really fast review of several techniques we can use to deal with long execution times and/or big data. From efficient code tips to the use of big analytics packages in just five minutes.

Keywords: Big data, Efficiency, Parallelization

Regression Spline Mixed Models for Analyzing EEG Data and Event-Related Potentials

Karen Nielsen

Department of Statistics, University of Michigan, Ann Arbor, USA
http://biosocialmethods.isr.umich.edu/

Abstract: Analysis of EEG data tends to be a nuanced, subjective process. For example, filtering is common, primarily to reduce noise, but a wide variety of filters are available with only heuristic (not theoretical) recommendations for use.

This work focuses on Event-Related Potentials (ERP), which generally involve waveforms with only one or a few oscillations. Since EEG readings consist of highly-correlated multi-channel readings, an ideal modeling approach should make use of this structure. Here, we will show how Regression Spline Mixed Models (RSMM) can combine the features of splines with a hierarchical framework to explore EEG data at any of the many levels that are collected and of interest to researchers.

While there are established protocols for working with EEG and ERP data in other software, R packages for this kind of data are more novel. This creates an opportunity to have innovative new approaches to working with neuronal waveform data included in the development of these packages, thus enabling use of these methods by non-statisticians. Here, we outline how existing functions in R can be combined to achieve a general approach to ERP data.

Keywords: basis sets, splines, hierarchical/mixed models, event-related potentials, EEG

rstats4ag.org, A website to help crop and weed scientists

Jens Carl Streibig

Plant and Environmental Sciences, University of Copenhagen, Denmark

http://plen.ku.dk/english/

Abstract: Being an open source, R is a programme and environment that has evolved over time. It can be accessed anywhere, also in developing countries with few resources for purchasing commercial programmes (e.g SAS, SPSS, Matlab). The dynamic of R makes the internet perfect for updating information of R functionality and R add-on packages and teaching R without costs. The purpose of the site, http://rstats4ag.org/, is simply to provide information and examples on how to use R to analyse statistical designs that are commonly used in crop and weed science experiments. The website is not meant to be a complete reference for all the capabilities of R, nor should it be used as a substitute for consulting well-trained statisticians. A majority of agricultural research uses a small subset of experimental designs, and thus, there is a high probability that the examples presented will provide a framework for analysis of many experiments. The codes for analysing common models, be it ANOVA, ANCOVA, linear or nonlinear regressions, and mixed models, will hopefully be useful to students and practitioners and will strengthen the statistics in the agricultural industries.

Keywords: ANOVA, ANCOVA, Regressions, mixed models

Teaching R graphics and visualization rhetoric

Richard Layton

Mechanical Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN, USA

http://www.rose-hulman.edu

Abstract: Approaches to teaching data visualization exist on a spectrum. At one end, learning objectives focus on R programming with little or no emphasis on rhetoric. At the other end, the focus is on visualization rhetoric or graphic design with little or no emphasis on software. Where a particular instructor stands on this spectrum tends to correlate with their program type. For example, statistics and computer science programs tend to focus on programming and statistical analysis, journalism and technical communication programs tend to focus on rhetoric, and business and engineering programs lie somewhere between.

I argue that being "somewhere between" is important for all students of data visualization. Data graphics are fundamentally about communication. R programming teaches "how"; visual rhetoric teaches "why". Both are needed at more than superficial levels if we are to fully prepare our students for the professions. In this talk, I describe the learning objectives and student activities in my data visualization course designed for the middle of the spectrum, with significant emphasis given to both why and how data visuals are created.

My goals are to prompt a discussion with other teachers and promote course design with significant emphasis on both R graphics and visualization rhetoric.

Keywords: R graphics, data visualization, teaching R, technical communication, visual rhetoric

Zombie Preparedness

Michael Höhle

Department of Mathematics, Stockholm University, Sweden
http://www.math.su.se

Abstract: It is just another day at the office, when suddenly the monitoring system signals a strange incoming aberration of zombie bites... This completely hypothetical use-case, extending the 'Preparedness 101' campaign by the Centers for Disease Control and Prevention (CDC) to a statistical setting, is about how the R package 'surveillance'. The package provides a statistical toolkit to detect, track and analyse outbreaks before, while and after they occur. Applied infectious disease epidemiology in the form of delay adjusted surveillance algorithms, nowcasting and back-projections is illustrated using a blitz of imaginary surveillance data, visualizations and models. At the end of this horror talk, you might find yourself better prepared for zombie suRvival.

Keywords: Infectious disease epidemiology, biostatistics, outbreak

Part IV

Posters

Posters

A Landsat Time Series Processing Chain using Parallel Computing, R and Open Source GIS software for Ecosystem Monitoring

Fabián Santos

University of Bonn

http://www.zfl.uni-bonn.de/

Abstract: Since the Distribution Policy of Landsat data enable the free download of the whole archive, the processing and analysis of large collections of Landsat images became a challenging task which demands efficient processing chains, not available yet in open source softwares. For this reason, we develop a sequence of R user-friendly scripts for organize the management, processing and extraction of land cover change patterns from a time series archive of the Landsat sensors TM, ETM+ and OLI-TIRS. For improve the computing time, we use the parallel computing approach and chose the best libraries and algorithms available from different open source geographic information systems for enable the geometric and radiometric correction standards, as well, the cloud, shadow and water masking, change detection and accuracy assessment. Our first prototype, gave us a map of the forest restoration since the 1984 to 2014 of a set of study areas distributed along an altitude gradient of the Amazon region of Ecuador. This processing chain constitutes a useful tool for ecosystem monitoring, evaluation of potential REDD+ projects, deforestation mapping, land grabbing, and other derived studies from Landsat time series analysis.

Keywords: Remote Sensing, Time Series Analysis, Parallel computing, Ecosystem monitoring

Adding a corporate identity to reproducible research

Thierry Onkelinx

Research Institute for Nature and Forest

https://www.inbo.be/en

Abstract: Markup languages like Markdown, HTML and LaTaX separate content and style. This distinction makes it fairly easy to apply a different style to a document. The knitr package facilitates to create reproducible documents by combining R code with the markup languages. The recent rmarkdown package converts R Markdown documents into a variety of formats including HTML, MS Word and PDF. We used these tools to create a package applying the style of the corporate in reproducible documents. The source code of the documents can be either LaTeX or Markdown. Dummy documents with various style items are added as vignettes to check the consistency with the corporate identity.

The main component of the package is a local texmf tree which contains the corporate identity of several types of documents (report, slides, poster). For Sweave files, only this part of the package is necessary. For R Markdown files, two additional components are necessary: Pandoc templates and R functions. The Pandoc templates select the appropriate LaTeX style and put the content of variables into the document. The R functions translate information in the YAML block of the Markdown file to the correct Pandoc template and required variables (title, author, cover image, language, ...).

Keywords: corporate identity, reproducible research, Markdown, LaTeX

Analysing student interaction data for identifying students at risk of failing

Jakub Kuzilek

Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
http://kmi.open.ac.uk/

Abstract: Our research aims at identification students at risk of failing the course at the Open University, UK. For the purpose of analysis we developed the system, which is using R as the platform for the at-risk student identification. The available data contain several demographical attributes such as gender, previous education, age, etc. and unique data about student interactions in the Virtual Learning Environment. These data is then processed using k-Nearest Neighbours (package FNN), CART decision tree (package rpart) and naïve Bayes classifier (package e1071). The information weather the student will pass/fail of submitting next assignment provided by each classifier is then combined using majority voting and final decision is made. We are delivering predictions for more than 25000 students every week. For evaluation of classification quality we are using precision and recall. Both measures varies in the time in course but the overall values are around 70%.

Keywords: Student data, At-risk prediction, Learning Analytics, Classification, Classification performance

Analysis and Prediction of Particulate Matter PM10 in Graz

Burger-Ringer Luzia

Department of Statistics, University of Technology, Graz, Austria
http://www.statistics.tugraz.at/

Abstract: In the region of Graz the threshold value (=50$\mu g/m^3$) of the average daily concentration PM10 is exceeded on more than 100 days of the year. This situation appears mainly within the six months October till March. So we investigated the influence of meteorological as well as anthropogenic factors based on data from the winter seasons 2002/03 to 2014/15. Exploratory data analysis shows that the emergence of wind and/or precipitation leads to lower values of PM10, whereas temperature inversion (lower temperature on the ground than above the ground) yields rather high values of PM10. This meteorological phenomenon can be observed up to 60% of the days in winter seasons and may be one explanation for extraordinary high PM10 values in and around Graz. However, the anthropogenic impact cannot be neglected, too. We will illustrate some scenarios which point out the influence of traffic and combustion processes.

Keywords: Environmetrics/Ecological Modeling/Meteorology, multiple linear regression, Forecasting of PM10, quality function, comparison observation-prediction

Analysis of massive data streams using R and AMIDST

Anders L. Madsen[†], **Antonio Salmeron**[‡]

[†]*Hugin Expert A/S, Aalborg, Denmark*
[‡]*Department of Mathematics, University of Almería, Spain*
http://www.hugin.com

Abstract: Today, omnipresent sensors are continuously providing streaming data on the environments in which they operate. Sources of streaming data with even a modest updating frequency can produce extremely large volumes of data, thereby making efficient and accurate data analysis and prediction difficult. Probabilistic graphical models (PGMs) provide a well-founded and principled approach for performing inference and belief updating in complex domains endowed with uncertainty. The on-going EU-FP7 research project AMIDST (Analysis of MassIve Data STreams, http://www.amidst.eu) is aimed at producing scalable methods able to handle massive data streams based on Bayesian networks technology. All of the developed methods are available through the AMIDST toolbox, implemented in Java 8. We show how the functionality of the AMIDST toolbox can be accessed from R. Available AMIDST objects include variables, distributions and Bayesian networks, as well as those devoted to inference and learning. The interaction between both platforms relies on the rJava package.

Keywords: HUGIN, AMIDST, data streams

Analysis of toxicology assays in R

Maxim Nazarov

Open Analytics, Antwerp, Belgium
http://www.openanalytics.eu/

Abstract: Toxicology testing is an indispensable part of the process of drug development. Genetic toxicity studies play important role in the safety assessment of a compound during preclinical stage. They are aimed at detecting whether compound induces DNA damage that can cause cancer or heritable defects.

We present a suite of R packages aimed to facilitate statistical analysis and reporting for commonly used genetic toxicity assays: invivo and invitro micronucleus tests and comet assay. The experiments usually follow a common hierarchical set-up, that allows to automate most of the analysis and reporting.

The statistical analyses implemented in the R packages follow the recent recommendations from the OECD guidelines for toxicity testing, and include fitting generalized linear models or mixed-effects models with appropriate hypothesis tests. A range of different plots can be created for data exploration. Additionally functionality to generate customized reports (including Word) is provided using R-markdown and pandoc with custom filters.

Keywords: biostatistics, toxicology, mixed-effects models, generalized linear models, reporting

Applications of Outlier Detection in R

Agnes Salanki

Department of Measurement and Information Systems, Budapest University of Technology and Economics, Budapest, Hungary

https://inf.mit.bme.hu/en

Abstract: According to the classical definition by Hawkins, outliers are observations deviating so much from the bulk of data that it is suspicious that they were generated by other mechanisms. Thus in several domains, like security (network intrusion detection) of finance (fraud detection), outlier identification and characterization is not only preparation of further statistical model building but an equivalent, individual step of data analysis.

Automatic outlier detection is supported in R, several implementations are available in packages like depth, fields, robustX, DMwR, etc.

The poster presents two use cases for applications of the above mentioned built-in R functions. First, outlier detection is used for finding performance anomalies in the behavior of our educational cloud functioning at our university since 2012. The cloud serves 300 students and is designed as a high-availability system (>99% availability), thus, identification of anomalous behavior causing further performance problems is vital for us. Secondly, conclusions of outlier detection performed on the PISA survey results are presented, including possible interpretations and open questions about school performance of students from individual countries.

Interpretation of results is supported with visualizations tailored to outlier detection.

Keywords: outlier detection, IT log analysis, PISA survey, visualization

Begin-R: Learning to use R within a National Statistics Institute

Amy Large

Office for National Statistics, UK

http://www.ons.gov.uk

Abstract: In April 2013, the UK Government Digital Service (GDS) released the Government Service Design Manual. This selection of documents includes guides on picking the right technological tools for jobs we want to do. Of open source software, the manual states that the Government has a level playing field between proprietary and open source software, and it should be actively considered when looking at software solutions. At the Office for National Statistics (ONS), the tools currently being used for statistical processing are predominantly traditional licence-based tools such as SAS and SPSS. With the freedom to use open source software, we are now in a position to make better use of R. But how do we bridge that gap between what we know and are comfortable with, and the new possibilities afforded to us with open source tools, both as an Office and as individuals? This paper will focus on my steps to ensure that I am embracing R and all it has to offer. I will also discuss how R is being used within ONS, and what steps are being taken to encourage analysts and developers to consider this as an alternative to the more traditionally used tools like SAS.

Keywords: Government, Software, Learning, Personal Development

Bias analyses in large observational studies: a non-trivial example from bovine medicine

Veit Zoche-Golob

Department of Bioprocess Engineering – Microbiology, Faculty 2, University of Applied Science and Arts, Hannover, Germany

http://f2.hs-hannover.de/organisation/labore/analytik/mikrobiologie/index.html

Abstract: The background of the analysis was an investigation of the association between the milk fat-protein ratio and the incidence of clinical mastitis in dairy cows. A mixed Poisson regression model for time-to-event data including repeated events and time-varying explanatory variables was fitted using the R package lme4. Because the recording of clinical mastitis might have been imperfect, a probabilistic analysis should be conducted to assess the direction and the magnitude of the misclassification bias on the conventional estimates. It was not feasible to refit the model several times due to its complexity. Therefore, data sets were simulated using the fitted model. A matrix adjustment method was used to simultaneously model the misclassification of the outcome (different across the number of previous events and the levels of the fat-protein ratio) and the misclassification of the number of previous events which depended on the misclassification of the outcome. Given the assumptions we made about the bias parameters and the methods we used, the conventional parameter estimates for fat-protein deviations were biased toward the null by 10-20% by the misclassification of clinical mastitis.

Keywords: veterinary epidemiology, bovine mastitis, observational study, sensitivity analysis, misclassification bias

Calculation of probabilities of the Mann-Whitney distribution using R

W. H. Moolman

Department of Statistics, Walter Sisulu University, Mthatha, South Africa
http://www.wsu.ac.za/waltersisulu/

Abstract: Mann and Whitney (1947) gave a recursive formula to calculate probabilities of this distribution. This formula is rather slow and not very useful from a computational point of view. Better computational algorithms that are based on the probability generating function of the distribution were suggested by a number of authors including Wilcoxon, Katti and Wilcox (1973) and Harding (1984). The R implementation of the three before mentioned algorithms will be shown. For sufficiently large sample sizes the normal approximation can be used. An R program for determining the sample sizes for which the normal approximation is accurate will also be presented.

Keywords: U-distribution, recursive, generating function, algorithms, normal approximation

circlize: circular visualization in R

Zuguang Gu

Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

http://ibios.dkfz.de/tbi/

Abstract: Circular layout is efficient to visualize multiple dimensional data. The pioneer software Circos [1] already makes great success for the circular visualization in many areas, especially for understanding huge amount of genomic data. Here we present the circlize [2] package which implements circular visualization in R as well as enhances available software. The package is based on the implementation of basic low-level graphics functions (e.g. drawing points and lines). Therefore, it is flexible to customize new types of graphics. In addition, with the seamless connection between data analysis and visualization in R, automatic procedures for generation of circular designs can be easily achieved. With the generality and simplicity of the package, circlize provides a basis on which high-level packages focusing on specific interests can be built.

We will demonstrate how to make close control on the circular layout, how to use low-level graphics function to build a complex circular plot and specific application e.g. on phylogenetics and genomics, Finally, we will demonstrate the customization of Chord Diagram which is useful to revealing complex relations in the data.

References [1] Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639-1645. [2] Gu Z, et al. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30:19

Keywords: Graphics, Circular visualization, Circos, Chord Diagram

Clinical-pharmacogenetic predictive models for MTX discontinuation in rheumatoid arthritis

Barbara Jenko

Institute of biochemistry, Faculty of Medicine, University of Ljubljana, Slovenia

Abstract: Objectives: Variability in genes involved in methotrexate (MTX) transport and target pathways was associated with MTX treatment outcome in rheumatoid arthritis (RA) patients. We investigated the effect of a large number of clinical and genetic factors on MTX discontinuation due to inefficacy and adverse events (AE). We developed and evaluated a prognostic index that could facilitate the translation of our results into clinic. Methods: In total 333 RA patients were genotyped for polymorphisms in folate and adenosine pathway and MTX transporters. Multivariable Cox models with LASSO penalisation was used to estimate prognostic factors and to construct the prognostic index. Its predictive capacity was evaluated with the cross-validated area under time dependent receiver operating characteristic curve (tAUC). Results: MTX dose, ABCG2 and ADORA2A were associated with discontinuation due to inefficacy, while RF or ACPA seropositivity, MTX dose, MTX monotherapy, SLC19A1, ABCG2, ADORA3 and TYMS were associated with discontinuation due to AE. Clinical-pharmacogenetic model of MTX discontinuation due to AE had better predictive ability than non-genetic model however the prediction was mostly worthless during the 2 years treatment time period. Conclusions: Application of those clinical-pharmacogenetic predictive models may support the future development of personalized MTX treatment in clinical practice.

Keywords: LASSO penalized regression, pharmacogenetic model, cross-validation

Collaborative statistics education through OpenIntro and GitHub

Andrew Bray[†], **Mine Çetinkaya-Rundel**[‡]

[†]*Department of Mathematics, Reed College, Portland, Oregon*
[‡]*Duke University*
http://academic.reed.edu/math/

Abstract: The userR! 2014 conference in Los Angeles featured an invited talk on OpenIntro, a project that develops free and open-source educational resources. The focus of the talk was on the development of the OpenIntro Statistics textbook as a collaborative and open-source enterprise and its parallels with the R project. At the end of the talk, by far the most common question by useR! participants was: is the textbook on GitHub?

In June 2014, the answer was no, but one year later, the answer is yes. The full textbook is now in a public GitHub repository as are the OpenIntro Labs, which teach statistics by analyzing real data in R. In this poster we would like to update conference participants on how educators at all levels can access, remix, and contribute content related to statistics education. We would also like to showcase a model for successful forks of educational materials, namely, R labs that have been translated into the R mosaic idiom.

Keywords: openintro, open source, github, education, collaborative

Convenient option settings management with the settings package

Mark van der Loo

Department of Methodology, Statistics Netherlands, The Hague

http://www.cbs.nl

Abstract: My recently published "settings" package is aimed to make option settings management in R more convenient. In particular, with this package one can:

- Define one's own option settings manager (with default settings) in a single call.
- Alter or request options like with "options()", but also reset all option values with a single call.
- Merging or altering option settings either globally or locally with ease (e.g. when writing functions with the "..." argument).

Besides that, the package offers a convenient way to reset "options()" or "par()" to their 'factory settings' in a single call. For example, calling

```
reset_par()
```

resets almost every graphical parameter to its default; exceptions are a few settings that typically have device-dependent defaults such as "mai" (margin size in inches) or "pin" (current plot dimensions in inches). A call to

```
reset_options()
```

resets all of R's options their defaults.

Keywords: Option settings management, Package building, R Programming

Converting to R in an FDA regulated industry: The trials and tribulations of deploying a GNU solution.

Robert Tell

Abbott Diagnostics, Abbott Laboratories

https://www.abbottdiagnostics.com/en-us/index.html

Abstract: Within the medical device industry, there is great reliance on the SAS programming language deriving from decades of SAS-based clinical trial data analysis. This proficiency has proliferated throughout industry over the years, creating both a wealth of programming experience and large libraries of code. This contributes to significant institutional inertia to use only SAS. R is a functionally equivalent language that offers to expand the toolset available for statistical programming. R allows for great flexibility due to its rich selection of open-source libraries, scalability, and ease of deployment. However, the adoption of R by a corporation within a regulated industry presented a series of challenges. These surrounded compliance both with the FDA and Abbott's own quality system. As part of these processes, there were the steps of vendor quality assurance, software quality assurance, and software code validation. Finally, there was the effort required to assure stakeholders that R could be used successfully for these tasks while maintaining regulatory and quality compliance. All of these proved significant challenges in extending analytical applications to R within the research and development organization at Abbott Diagnostics. Herein, the process will be outlined and a roadmap for others might be furnished in developing compliant R applications.

Keywords: FDA, SAS, Validation, 21CFR11, Medical Devices

Creating a Shiny dashboard for a legacy integrated library system

Matti Lassila

Jyväskylä University Library, Finland

https://kirjasto.jyu.fi/?set_language=en

Abstract: The integrated library system (ILS) has been traditionally the backbone of all library operations, including acquisition of the resources, cataloguing and collection management. Therefore, a wealth of information is being stored to the ILS and is potentially available for analysis.

Unfortunately, the built-in reporting capabilities of the ILS are usually very limited. Sometimes, these limitations can be circumvented using external database query tool if the ILS vendor permits direct SQL-access to the relational database powering the ILS.

In our case we were using MS Access as a primary reporting interface to the Oracle 11g database of the ILS. At the request of non-technical staff members systems librarian created MS Access queries on ad-hoc basis. This workflow was time consuming and because of the manual nature of the process, it was impossible to utilize real time information, such as book hold statuses or transaction logs.

Using existing SQL queries as a starting point, we created a Shiny web app to automatize and greatly improve our reporting process. In addition to the Shiny the key building blocks have been ROracle and scheduleR, which we are using as a lightweight Extract-Transform-Load (ETL) tool.

Keywords: dashboard, reporting, legacy systems, integrated library system, academic library

Creating an half-automated data preparation workflow for Codelink Bioarrays with R and Bioconductor

Anita Höland

Institute for Medical Informatics, Justus-Liebig-University, Giessen, Germany
http://www.uni-giessen.de/cms/fbz/fb11/institute/imi/ag/agms

Abstract: Microarray analysis has been developed to identify the expression levels of a large number of genes simultaneously. R and the Bioconductor platform are then widely used to conduct the necessary data preparation and analysis steps to identify emerging patterns of gene expression. Packages like Limma1were developed in the early beginnings of R and the Bioconductor projects supplies a large variety of R-packages for data analysis for different kinds of microarrays. The package Codelink2, which we use, was developed to preprocess and analyze Codelink Bioarrays (GE Healthcare). To use these packages and to prepare the data for analysis a certain amount of knowledge about the required operations and the usage of R is needed. To ease and speed up this process, for user without knowledge or R, we are developing a workflow in R which then only requires the user to input certain parameters to produces the output in graphs and tables. The workflow includes functions from existing packages and newly implemented functions developed for this purpose. To broaden the range of applicable studies the workflow can be adapted for supervised (case-control studies) and unsupervised (cluster analysis) study designs.

Keywords: microarray analysis, transcriptome analysis, data preparation, Codelink

DDIR and dlcm : integrated environment for social research data analysis

Yasuto Nakano

School of Sociology, Kwansei Gakuin University, Nishinomiya, Japan
http://www.kwansei.ac.jp/

Abstract: The purpose of this presentation is to propose an environment for socialresearch data and its analysis.

A R package DDIR and an IDE dlcm, which utilize social research informations in DDI format, offer you integrated environments for social research data. DDI(Data Documentation Initiative) is a XML protocol to describe informations related to social research including questionnaire, research data, meta data and summary of results. There are several international research projects which use this protocol as a standard format. ICPSR(Interuniversity Consortium for Political and Social Research), one of the biggest data archive for social research data, encourages data depositors to generate documentation that conforms with DDI.

In R environment, there is no standard data format for social research data . In many case, we have to prepare numerical data and label or factor informations separately. If we use DDI file as a data file with DDIR in R, only one DDI file is needed to be prepared. DDI file could be a standard data format of social research data in R environment, just same as 'sav' file in SPSS. DDIR realizes integrated social research analysis environment with R, and ensures it as a reproducible research.

Keywords: DDI, xml, data format, social research data, reproducible research

DDNAA: Decision Support System for Differential Diagnosis of Nontraumatic Acute Abdomen

Gokmen Zararsiz[†], Hizir Yakup Akyildiz[†], Dincer Goksuluk[‡], Selcuk Korkmaz[‡], Ahmet Ozturk[†], Sevilay Karahan[‡]and Eda Karaismailoglu[‡]

[†]*Dept. of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey*
[‡]*Dept. of Biostatistics, Hacettepe University, Ankara, Turkey*

http://tip.erciyes.edu.tr

Abstract: A quick evaluation is essential for patients with acute abdominal pain. It is crucial to differentiate between surgical and nonsurgical pathology to prevent mortality and morbidity. Practical and accurate tests have importance in this differentiation. Recently, D-dimer level is found to be an important marker in this diagnosis and obviously outperforms leukocyte count, which is widely used for diagnosis of certain cases. Here, we built DDNAA, a user-friendly shiny application, to assist physicians in their decisions to diagnose patients with acute abdomen. An experimental study is conducted and 28 statistical learning approached were assessed for this purpose by combining leukocyte count and D-dimer levels in order to make an increase in the diagnostic accuracies. DDNAA web-tool includes the best performed algorithms naïve Bayes, robust quadratic discriminant analysis, k-nearest neighbors, bagged k-nearest neighbors and bagged support vector machines that provided an increase in diagnostic accuracies up to 8.93% and 17.86%, comparing to D-dimer level and leukocyte count, respectively. DDNAA shiny application is available at http://www.biosoft.hacettepe.edu.tr/DDNAA/.

Keywords: Abdominal pain, D-dimer level, Decision support system, Nontraumatic acute abdomen, Statistical learning

Density Legends

Jason Waddell

Open Analytics

http://www.openanalytics.eu/

Abstract: Traditional color legends present a missed opportunity for gaining added insight into the color variable. The densityLegend() function introduces functionality that combines the legend with a color-partitioned density trace, for visualizing the distribution of the color variable. In addition to legends, color-paritioned density traces introduce a range of unique visualizations.

We present a package for easy integration of density legends into base R plots, with a planned ggplot2 adaptation.

Keywords: visualization, legends, density plots, color

Directional Multiple-Output Quantile Regression in R

Pavel Bocek

Department of Stochastic Informatics, UTIA AVCR, the Czech Republic
http://www.utia.cas.cz

Abstract: The presentation introduces an R package for performing two recent directional multiple-output quantile regression methods generalizing Koenker's quantile regression to the case of multivariate responses. It starts with a necessary but brief theoretical introduction, continues with a brief description of the R package and its functionality, and concludes with a carefully designed set of practical illustrative examples how the package can be used to solve the parametric optimization problems behind both of the directional multiple-output quantile regression approaches, to evaluate the resulting regression quantile contours or their cuts, and to compute various meaningful inferential statistics. The applications include locally constant multiple-output quantile regression and the computation of halfspace depth contours in two to six dimensional spaces, among others. In summary, the R package finally makes the two promising multiple-output quantile regression methods freely available to the statistical public.

Keywords: multivariate quantile, quantile regression, multiple-output regression, data depth, linear optimization

Discrete event simulation and microsimulation.

Andreas Karlsson

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

http://ki.se/en/meb

Abstract: Discrete event simulation (DES) is a powerful technique for simulating complex systems. Surprisingly, there are few options for DES in R. Microsimulation is a modelling technique that operates at the level of individuals. We required a flexible DES framework for microsimulation of prostate cancer screening.

For the microsimulation package (https://github.com/mclements/microsimulation), we provide a pedagogic DES implementation with R5 classes. However, the R5 implementation and a C++ process-oriented DES scale poorly for 10^6 individuals. For speed and flexible simulation specification, we use a C++ event-oriented DES library as the simulation core. A natural workflow uses R for pre- and post-processing, with complex data structures passed between R and C++ using Rcpp. We have developed several tools to support the microsimulation. For variance reduction, we use common random numbers, with stream manipulation at the C++ level. We also use C++ reports to substantially reduce the post-processing burden in R. We demonstrate the package by predicting the cost-effectiveness of prostate cancer screening for different screening scenarios.

The combination of Rcpp and C++ allows for a fast DES framework with easy data management and analysis in R.

Keywords: discrete event simulation, microsimulation, Rcpp, common random numbers, prostate cancer screening

Disease mapping and ecological regression for Belgium Shiny application

Tom De Smedt

KULeuven

http://www.kuleuven.be

Abstract: Disease mapping is one of the most widely used techniques in epidemiology. In disease mapping we divide the area of interest in different subareas, and we display the disease rate for each of these subareas, potentially allowing us to identify anomalous subareas. In order to be able to correctly compare subareas when the disease under investigation is statistically rare, modern disease mapping methods use different kinds of spatial smoothing methods. In ecological regression we also have an exposure variable for each subarea, which allows us to look at the relationship between the disease and exposure on the subareal level. We have developed a Shiny application, where the user can upload the disease data (and the exposure data) for each subarea. The application can then correctly map the data, where the user can choose between different smoothing methods (no smoothing, Bayesian hierarchical smoothing methods, spline smoothing). For ecological regression, both exposure and disease data are uploaded, and the user can choose between different methods to investigate the relationship. Given the large rise in available datasets, this application allows epidemiologists to quickly evaluate novel data using an easy-to-use and lightweight platform.

Keywords: epidemiology, GIS, Shiny

Documents Clustering in R

Sonya Abbas

E-Government unit, NUIG, GALWAY, IRELAND

https://www.insight-centre.org/

Abstract: The pressure of evaluating and improving the government's actions plans is driven by a combination of factors. These factors include the difficulties of discovering the similarities between actions plans of different countries as this helps learning from similar experience and the need to evaluate the actions plans based on the challenges and commitments they address. Documents clustering have been used as an approach in order to solve this problem. However, This wasn't straight forward as we face challenges related to data preparation such as dividing the actions plans into challenges and commitments, preparation such as filtering the features and visualization. Data Science steps have been followed from data collection, cleaning, preparation, analyzing, visualization and interpretation. We use R for implementing this. Kmeans and hierarchical clustering has been applied and different visualization ways has been presented using different packages like ade4, ggplot2, ellipse, HSAUR and flexclust. As an evaluation, we suggest internal and external quality measures like entropy, F measure and overall similarities and imply it using R packages.

Keywords: actions plans, documents clustering, kmeans, hierarchical clustering

Dose reconstruction based on questionnaires with taking into account human factor uncertainty

Konstantin Chizhov

Laboratory of radiation-health studies, Burnasyan Federal Medical Biophysical Center, Moscow, Russia

http://fmbcfmba.ru/

Abstract: The article shows developed methodology for evaluating the reliability of dose reconstruction taking into account human factor uncertainty. In the absence of radiation monitoring, analysis of doses is based on questionnaires - recorded on paper stories about people activities. This situation is very typical for accidents when there are not enough dosimetry equipment. Thus, there is a chance that the questionnaire will contain some information that does not correspond to reality, and some information will be absent or be invented. In our study we have checked questionnaires of 28 emergency workers. We used two questionnaires for each worker - one was filled in the first year after the accident, and the second - after 20 years. For comparison we transformed questionnaires to a matrix of elemental fragments with a set of parameters: dose rate, coordinates, residence time, protective factor. Dose rate in every location was calculated using RADRUE method by interpolation of measured values. Through this analysis, we calculated contribution of the human factor in uncertainty of doses and find places where workers reported incorrect information.

Keywords: Dose reconstruction, Human factor uncertainty, Multivariate Analysis, Fragmentation, Analysis of questionnaires

Early Detection of Long Term Evaluation Criteria in Online Controlled Experiments

Yoni Schamroth

Perion Networks

http://perion.com

Abstract: Controlled Experimentation has been universaly adopted by the online world as an essential tool in aiding in the decision making process and (maybe new sentence) has been widely recognized as a successful scientific method for establishing causality. Frequently referred to A/B testing or multivariate testing, controlled experiments provide a relatively straightforward method for quickly discovering the expected impacts of new features or strategies. One of the main challenges involved in setting up an experiment is deciding upon the OEC, or overall evaluation criteria. In this paper, we demonstrate the importance of choosing a metric that (focuses on, captures, emphasizes) long term effects. Such metrics include measures such as life-span or lifetime value. We present motivating examples where failure to focus on the long term effect may result in an incorrect conclusion. Finaly we present an innovative methodology for early detection of lifetime differences between test groups.

Keywords: Multivariate Testing , Resampling , A/B Testing, Life Time Value

Easily access and explore your Hadoop Big Data with visualisations and R/TERR jobs

Ana Costa e Silva

Tibco Software Inc.

http://www.tibco.com

Abstract: TIBCO Spotfire®- Hadoop integration points can be grouped into two categories: TIBCO Spotfire native data connectors and TERR (TIBCO's enterprise platform for the R language)-Hadoop integration. Both provide an extensive set of analytic features and security options.

Spotfire Hadoop connections can be quickly configured into analytic workflows, dashboards, or reports, which can then be shared, reused, and consumed across organizations. KPIs based on Hadoop data can be pushed to virtually any user device via HTML. Extensive geo analytic support within Spotfire makes it easy to generate insights from geographical data.

In this session, we will explain and demo the powerful combination of Spotfire, TERR, and Hadoop and how it enables deeper, more valuable analysis of Hadoop data. We will demonstrate how with it the business user can have an easy to use front-end from which to:

1. visualise big data interactively with surprising performance,
2. with just a few clicks, deploy map-reduce jobs, which run R code in the TERR engines installed in the Hadoop data nodes,
3. with again just a few clicks, launch H2O jobs when running calculations on all data-nodes at once,
4. consume the result of calculations, e.g. predictive models, and deploy them in real-time.

Keywords: Big and fast data, TERR and R, Hadoop, Visualisation, Map-reduce and H2O

easyROC: an interactive web-tool for ROC analysis

Dincer Goksuluk[†], Selcuk Korkmaz[†], Gokmen Zararsiz[‡], Sevilay Karahan[†], A. Ergun Karaagaoglu[†]

[†]*Dept. of Biostatistics, Hacettepe University, Ankara, Turkey*
[‡]*Dept. of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey*
http://www.biostatistics.hacettepe.edu.tr

Abstract: ROC analysis is a fundamental tool for evaluating the performance of a marker in number of research areas, e.g., biomedical, bioinformatics, engineering etc., and is frequently used for discriminating cases from controls. There are number of analysis tools guiding researchers through their analysis. Some of these tools are commercial and provide basic methods for ROC analysis while some others come up with advanced analysis techniques and command based user interface, such as R programming. R programming includes comprehensive tools for ROC analysis, however, using command based interface might be challenging and time consuming when a quick and reliable evaluation is desired especially for non-R users, physicians etc. Hence, a quick, comprehensive, free and easy-to-use analysis tool is demanded. For this purpose, we developed a user-friendly web-tool which is based on R language. This tool provides ROC statistics, graphical tools, optimal cut point calculation and comparison of several markers to support researcher in their decision without writing R codes. easyROC can be used via any device with internet connection free from device configuration and operating system. The web interface of easyROC is constructed with an R package shiny. This tool is freely available through www.biosoft.hacettepe.edu.tr/tools.html.

Keywords: receiver operating curve, optimal cut point, diagnostic test, web tool, discrimination

Enlighten the past: The R package Luminescence - signal, statistics and dating of environmental dynamics -

Sebastian Kreutzer

IRAMAT-CRP2A, Université Bordeaux Montaigne, France

http://www.iramat-crp2a.cnrs.fr

Abstract: Earth surface processes decisively shape our planet. To decipher the timing and rates of Earth surface pro- cesses throughout the last 250,000 years, one numerical dating method has reached paramount importance: Luminescence dating. This method provides robust numerical data on environmental changes by measuring the luminescence signal of minerals, which is reset by daylight exposure or heating and has the advantage of using nearly ubiquitously available mineral grains of quartz or feldspar. During the last decades more and more luminescence-based ages have been requested and the method has been considerably enhanced. However, an increasing methodological complexity demands for a flexible and scalable software solution for data analysis. The presented R package Luminescence is designed as a toolbox intending to provide customised solutions for a variety of requirements, e.g. measurement data import, statistical analysis, graphical output. The used algorithms and statistical treatments are always transparent and the user maintains in control of combining and adjusting algorithms by taking advantage of the wide range of functions available in R. Our contribution summarises the concept of the R package Luminescence and focuses on some conceptional aspects and selected practical examples.

Keywords: Luminescence Dating, Geosciences, Dosimetry, Signal Analysis

Extending the Quasi-Symmetry Model: Quasi-Symmetry Model with n Degree Symmetry

Tan Teck Kiang

Institute for Adult Learning, Workforce Development Agency, Sngapore
http://www.ial.edu.sg/index.aspx?id=58

Abstract: The quasi-symmetry model (QS) is one of the doubly classified non-standard log-linear models commonly used in social sciences in examining relationship of cells within a square table. The main characteristic of QS is that it exhibits symmetry in odds ratios for off-diagonal cells. A new proposed model, quasi-symmetry model with n degree symmetry, QS(n), relaxes this symmetry odds ratios assumption to a general QS with varying degree of symmetry. When the degree of symmetry at the lowest level with n=1, the QS(1) model, only those cells closest to the diagonal are in symmetry and those further away are freely estimated. The number of cells in symmetry goes up when the degree of symmetry n increases, thus formed a series of QS models with symmetry degree n. The QS(n) models are fitted using generalized linear model. Package R function gnm is used to fit the QS(n) model. Using a survey data that aims to examine the association between literacy skills and problem solving skills, the results show that QS(1) models fit better than QS model in explaining the association between the two skills, indicating the incremental information content and usefulness of QS(n) model.

Keywords: log-linear model, doubly classified model, categorical data analysis, package gnm

From data.frames to data.tables: optimization strategies for analyzing petabytes of cancer data

Malene Juul , Johanna Bertl, Qianyun Guo, Asger Hobolth, Jakob Skou Pedersen

Department of Molecular Medicine (MOMA), Aarhus University Hospital, Denmark
http://moma.dk

Abstract: In the age of big data, efficient data handling and analysis are challenging tasks. R gives access to a comprehensive and well maintained set of tools valuable for doing statistics and data analysis. However, these advanced functionalities often come at the price of calculation speed.

In cancer genomics, data sets with billions of data points are routinely produced. In this work we analyze 2,500 whole genome DNA sequences, each consisting of approximately three billion data points. In this setting the bottlenecks are efficient and accurate analysis methods.

Here, we are interested in determining the distribution of the sum of independent discrete stochastic variables using a dynamic programming approach for mathematical convolution. The chosen granularity of the discretization is a trade-off between calculation accuracy and speed efficiency. I will cover a variety of the code optimization strategies applied in the R implementation, e.g. the relatively simple changes needed to extend the primary R data structure of data.frames to the enhanced version of data.tables.

Keywords: optimization, data.tables, genomics, big data

Goodness-of-fit tests for the Exponential and the Weibull distributions

Meryam Krit

Open Analytics

http://www.openanalytics.eu

Abstract: Although the Weibull distribution is widely used in many areas such as: engineering, biomedical sciences, economics, reliability, etc. The checking of its relevance for a given data set is not always done or done by elementary techniques such as Weibull plots. There exist more sophisticated techniques which aim to determine if a given model is adapted to a given data set; the goodness-of-fit (GOF) tests. Many GOF tests for the Weibull distribution have been developed over the years, but there is no consensus on the most efficient ones. The aim of the talk is to present the R package EWGoF that gives an overview of up-to-date GOF tests for the two-parameter Weibull and the Exponential distributions. It contains a large number of the GOF tests for the Exponential and Weibull distributions classified into families: the tests based on the empirical distribution function, the tests based on the probability plot, the tests based on the normalised spacings, the tests based on the Laplace transform and the likelihood based tests, ...

An illustrative application of the GOF tests to real data sets is carried out at the end of the talk.

Keywords: Goodness-of-fit , Weibull distribution, Exponential distribution, likelihood based tests

Hough: analytic curves detection using the Hough transform

Pavel Kulmon, Jana Noskova, David Mraz

Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Czech Republic

http://www.fsv.cvut.cz/index.php.en

Abstract: The Hough transform is a feature extraction technique and its purpose is to find imperfect instances of objects within a certain class of shapes. The transformation has been successfully used in several areas such as computer vision, image analysis and last but not least, photogrammetry and remote sensing. The foundation of this work is built upon the book „Introduction to image processing using R: learning by examples “ written by A.C. Frery and T. Perciano, where were outlined options for working with the digital images using the R-project. Our main aim is to develope the new R package Hough. In this package algorithms for non-user image evaluation will be implemented. Now the package contains the Hough transformation for a line detection using the accumulator. We also want to incorporate other Hough techniques like the recognition of more advanced analytic curves. The preparation of the non-user image evaluation consists of various methods such as the image grayscaling, the thresholding or the histogram estimation. The transfer from the gray scaled image to binary is realised by the aproximation of the derivatives, which were computed by the Sobel operator. The new R package Hough will be the collection of all mentioned techniques.

Keywords: computer vision, convolution, Hough transform, edge detection

Hyperspectral Data Analysis in R: The new hsdar-package

Lukas W. Lehnert

Department of Gography, Philipps-University Marburg, Germany
http://www.uni-marburg.de/fb19

Abstract: An R software package is introduced which focuses on the processing, analysis and simulation of hyperspectral (remote sensing) data. The package provides a new class (Speclib) to handle large hyperspectral datasets and the respective functions to create Speclibs from various types of datasets such as e.g., raster data or point measurements taken with a field spectrometer. Additionally, the package includes functions for pre-processing of hyperspectral datasets and gives access to the vegetation reflectance simulation models PROSPECT and PROSAIL. The functionality of the package to analyze hyperspectral datasets encompasses a huge range of common methods in remote sensing, such as the transformation of reflectance spectra using continuum removal, linear spectral unmixing, the calculation of normalized ratio indices and over 90 different hyperspectral vegetation indices. Additionally, a direct access to multivariate analysis tools such as generalized linear models and machine learning algorithms via the caret-package is provided. The contribution shows a subset of available methods which are demonstrated by the analysis of 3D hyperspectral data taken to investigate effects of CO2 enrichment on grassland vegetation.

Keywords: Hyperspectral remote sensing, Canopy reflectance simulation, Continuum removal, Linear spectral unmixing

KFAS: an R package for exponential family state space modelling

Jouni Helske

Department of Mathematics and statistics, University of Jyväskylä, Finland
https://www.jyu.fi/maths/en

Abstract: State space modelling is an efficient and flexible method for statistical inference of broad class of time series and other data. Structural time series, ARIMA models, and generalized linear mixed models are just some examples of models which can be written as a state space model. Standard methods are often restricted to Gaussian observations due to their analytical tractability. I introduce an R package KFAS (Kalman Filtering And Smoothing), which can be used for state space modelling with the observations from exponential family, namely Gaussian, Poisson, binomial, negative binomial and gamma distributions. After introducing the basic theory behind the state space models and the main features of KFAS, an illustrative example for forecasting alcohol related deaths in Finland is presented.

Keywords: state space models, time series, R package, non-gaussian observations, forecasting

Large-scale multinomial regression: Modelling the mutation rates in whole-genome cancer data

Johanna Bertl

Department of Molecular Medicine, Aarhus University, Denmark
http://www.moma.dk/

Abstract: Understanding the mutational process in cancer cells is crucial to distinguish driver mutations, responsible for the initiation and progress of cancer, from passenger mutations. The heterogeneity of the process on various levels makes this a challenging question: whole-genome analyses have shown that the mutation pattern differs fundamentally between different cancer types, but also between patients and along the genome.

Here, we analyse whole-genome DNA sequences of tumor and healthy tissue of 505 patients with 14 different cancer types (Fredriksson et al., Nature Genetics, 2014). We model the probabilities of different types of mutations at each position on the genome by multinomial regression. Explanatory variables capture local genomic characteristics like the local base composition, the functional relevance of the region and epigenetic factors.

The enourmous dataset creates two different computational challenges: First, with the 3 billion basepairs of the human genome, n is very large. This requires to save and analyse the data in a compact format, even at the cost of loosing information. Second, including interactions between the patient ID and genomic variables considerably increases the number of parameters to estimate and thereby creates convergence problems. We approach these challenges using existing R-packages and our own developments.

Keywords: cancer genomics, large n, multinomial regression, numerical optimization, maximum likelihood estimation

Learning Graphical Models for Parameter Tuning

Marco Chiarandini

Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

http://www.imada.sdu.dk

Abstract: Recent research in the field of heuristic algorithms for nonlinear optimization has focused on methods to fine tune the main heuristic decision that are embedded in these algorithms. These decisions can be represented as categorical and numerical parameters. We designed and build a method to find the best setting of parameters based on graphical models and Bayesian learning. Each parameter is modeled by a node of the network and dependencies assumed a priori by arcs. Nodes have associated a local probability distribution, that for continuous parameters is given by Gaussian linear regression. Learning is achieved by a combination of importance sampling and Bayesian calculus. We implemented the method in R building on the package deal. Every data point corresponds to a run of the algorithms to tune and it may be computationally expensive, therefore we used Rmpi to execute the algorithms in parallel in a distributed environment. The results on two test cases, the traveling salesman problem and a nonlinear continuous optimization case derived from least median of squares, show that the method achieves competitive results with respect to state-of-the-art automatic tuning systems. More extensive testing is needed.

Keywords: Graphical models, Automatic parameter tuning, Bayesian learning, Optimization algorithms

Logr: An R package for logging in the R idiom

Davor Cubranic and Jenny Bryan

Department of Statistics, University of British Columbia, Canada

http://www.stat.ubc.ca

Abstract: *Logr* implements a logging framework that users R's existing messaging functionality, and builds upon it an API that is simple to use. Because it uses the same underpinnings, *logr* can capture output generated by `message`s and `warning`s, making it easy to adopt even in a mature codebase.

Most logging packages today seem to copy Java log4j API. It's important to remember that Log4j originated as the logging code for the Apache web server. As such, it was designed for use in large, long-running, and complex applications that contain many subsystems and potentially produce many output events per second.

Logr instead targets a lighter-weight usage scenario – arguably more likely in an R codebase – of a relatively short, often interactive script or command-line utility. To this purpose, *logr* provides a minimal API with sensible defaults that requires little effort to use in code.

Still, the API is powerful enough to allow multiple logging destinations, each with its own level of detail. This makes it simple, for instance, for a script to provide informational progress messages to the user, while recording detailed output in a log file.

Keywords: logging framework, programming, printing and error handling

Mapping the distribution of marine birds in the Northeast and Mid-Atlantic using a space-time double-hurdle model

Earvin Balderama

Department of Mathematics & Statistics, Loyola University Chicago, Chicago, Illinois, USA

http://www.luc.edu/math/

Abstract: The spatial distribution and relative abundance of marine birds along the US Northeast and Mid-Atlantic coastlines are of special interest to ocean planners. However, marine bird count data often exhibits excessive zero-inflation and extreme over-dispersion. Our modelling effort incorporates a spatial-temporal double-hurdle model specifically tailored to look at extreme abundances, which is especially important for assessing potential risks of offshore activities to seaducks and other highly aggregative species. We discuss several distributional forms of each component of the model, including negative binomial, log-normal, and a generalized Pareto distribution to handle the extreme right tails. Spatial heterogeneity is modelled using a conditional autoregressive (CAR) prior, and a Fourier basis was used for seasonal variation. Model parameters are estimated in a Bayesian hierarchical framework, using an MCMC algorithm with auto-tune parameters, all written and run in R. We demonstrate our model by creating monthly predictive maps (using ggmap) that show areas of high probability of aggregation and persistence for each of over fifty "high-priority" species as listed by Marine-life Data Analysis Team (MDAT) in collaboration with the Northeast Regional Ocean Council (NROC). A Shiny (R-Studio) app is currently being developed for quick reference of a desired space-time-species map.

Keywords: spatial statistics, mapping, extreme count distribution, Bayesian MCMC methods, zero-inflation models

Measuring dissimilarities between point patterns using R

Jonatan A. González)

Department of Mathematics, University Jaume I, Castellón de la Plana, Spain

http://www.uji.es/UK/departaments/mat/estructura/personal/e@/22752/?p_url=/UK/departaments/mat/estructura/personal&p_item=22752&p_per_id=423065

Abstract: Point processes are random collections of points falling in some space, this concept is used in order to describe a huge set of natural phenomena in a wide variety of applications.

Our interest concerns the spatial point processes, where each point represents the location of some object or event, such as a tree or sighting of a species. The classical model for a point processes is the Poisson process, where the numbers of points in any disjoint sets are independent random variables. The Poisson process is a natural null model in the absence of clustering or inhibition.

In order to differentiate between individuals based on their patterns, it is necessary the definition of a distance between two point patterns.

The purpose of this work is to outline several types of distances (and non-metric measures of dissimilarity) between two point patterns, and . We aim to implement dissimilarity measures based on functional or scalar descriptors of point processes, including estimators of first and second moments of the processes, or classical test statistics based on these moments. Finally we perform a simulation study and a real data analysis through functions and packages in R.

Keywords: Classification, K-function, Multidimensional scaling, Point patterns, Spike-time distance

MLSeq: Machine Learning Interface for RNA-Seq Data

Gokmen Zararsiz[†], Dincer Goksuluk[‡], Selcuk Korkmaz[‡], Vahap Eldem[§], Izzet Parug Duru[♭], Turgay Unver[♮], Ahmet Ozturk[†]

[†]*Dept. of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey*
[‡]*Dept. of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey*
[§]*Dept. of Biology, Faculty of Science, Istanbul University, Istanbul, Turkey*
[♭]*Dept. of Physics, Faculty of Science, Marmara University, Istanbul, Turkey*
[♮]*Dept. of Biology, Faculty of Science, Cankiri University, Cankiri, Turkey*
http://tip.erciyes.edu.tr/

Abstract: With the recent developments in molecular biology, it is feasible to measure the expression levels of thousands of genes simultaneously. Using this information, one major task is the gene-expression based classification. With the use of microarray data, numerous classification algorithms are developed and adapted for this type of classification. RNA-Seq is a recent technology, which uses the capabilities of next-generation sequencing technologies. It has some major advantages over microarrays such as providing less noisy data and detecting novel transcripts and isoforms. These advantages can also affect the performance of classification algorithms. Working with less noisy data can improve the predictive performance of classification algorithms. Further, novel transcripts may be a biomarker in related disease or phenotype. MLSeq package includes several classification and feature selection algorithms, also normalization and transformation approaches for RNA-Seq classification. MLSeq is available at http://www.bioconductor.org/packages/release/bioc/html/MLSeq.html

Keywords: Machine learning, Next-generation sequencing, Transcriptomics, RNA-Seq classification

Modeling the oxygen uptake kinetics during exercise testing of patients with chronic obstructive pulmonary diseases using nonlinear mixed models

Florent Baty

Department of Pulmonary Medicine, Cantonal Hospital St. Gallen, Switzerland
http://www.pneumologie.kssg.ch/

Abstract: Six-minute walk tests (6MWT) are common examinations performed on lung disease patients. Oxygen uptake (VO2) kinetics during 6MWT typically follows 3 phases that can be modelled by nonlinear regression. Simultaneous modeling of multiple kinetics requires nonlinear mixed models which, to our knowledge, have not yet been fitted in practice.

The aim is to describe functionality of the R package medrc that extends the nlme package framework of fitting nonlinear mixed models with a user-friendly interface, and demonstrate its usefulness in pulmonary medicine.

6MWT VO2 kinetics were measured on 61 patients with chronic obstructive pulmonary disease classified into 3 severity stages. A 6-parameter nonlinear regression model was defined and fitted to the set of kinetics using the function medrm(), allowing for automated fitting of a single joint nonlinear mixed model on multiple curves by extending the functionality of nlme().

All kinetics phases were incorporated within a single mixed model including fixed factors stratified into 3 clusters (disease stages), together with patient-specific random effects. Significant between-stage differences were found regarding maximum VO2 during exercise testing, inflection point and oxygen level at recovery.

medrc provides a comprehensive framework for the parametrisation and inference for hierarchical nonlinear mixed-effects regression models in various biomedical applications.

Keywords: nonlinear mixed effects models, exercise testing oxygen uptake kinetics, chronic obstructive pulmonary disease, medrc, nlme

Mutual information Implementation with R for Continuous Variables

Joe Suzuki

Osaka University

http://www.math.sci.osaka-u.ac.jp/eng/index.html

Abstract: When we deal with data analysis with R, computing mutual information (MI) is needed very often. For two discrete variabes, it is easy to compute the MI (one can construct a function with R easily). On the other hand, if they are Gaussian, it will be easy as well because only the correlation coefficient should be estimated. In this work, we condider the most general case in which its density function may not exist. The estimator is the difference of the BIC values when they are dependent and independent. The same principle works for discrete and continous cases. The estimated MI is strongly consistent, and is almost surely negative if and only if the two variables are independent.

The implementation is based on Ryabko's measure (2009). Given samples of size n, the estimation completes in O(n log n). In the presentation, we show the source program and several experimental results using the package.

Finally, we show the same principle is useful in learning a graphical model structure given examples if we extend the estimator of MI of X, Y to that of conditional MI of X, Y w.r.t. Z that can detect conditional independence of X, Y given Z.

Keywords: mutual information, continuous variables, R package, graphical models

Package webs: Reproducible results from raw data

Kirill Müller

IVT, ETH Zurich

http://www.ivt.ethz.ch

Abstract: For reproducible research, it is crucial to be able to generate all results from original raw data. By automating the process, it is possible to easily verify reproducibility at any stage during the analysis. Automation also allows easy recreation of the entire analysis based on modified inputs or model assumptions. However, rerunning the entire analysis starting from raw data soon becomes too time-consuming for interactive use. Caching intermediate results alleviates this problem but requires a robust mechanism for cache invalidation. R packages are a suitable container for statistical analyses: They can store data, code, and documentation. Recent efforts have considerably simplified the packaging process. This poster presents an approach to conduct a statistical analysis by creating a "package web" – interdependent packages where each serves a dedicated purpose (e.g., holding raw data, munging data, input validation, modeling, analysis, reporting, ...). Package dependencies define the data flow for the entire analysis. The "rpkgweb" companion package tracks which downstream packages need to be rebuilt if a package changes, and builds independent packages in parallel. Reproducibility can be monitored continuously with minimal effort, yet the modular structure permits interactive work.

Keywords: Reproducibility, Automation, Interactive

PASWR2 library for teaching with R

Ana F. Militino

Universidad Pública de Navarra

Abstract: PASWR2 is the second version of PASWR, acronym of Probability and Statistics with R. This package contains data sets, functions and scripts created for solving exercises, problems and theoretical questions of the book entitled with the same name. Its goal is to teach Statistics at an intermediate level using both mathematical classical tools, and R. Traditional scripts following theoretical formulae and programmed commands to illustrate them are presented.

Keywords: PASWR2, Teaching, Text book

R users all around the world

Gergely Daróczi

Easystats Ltd, United Kingdom

http://rapporter.net

Abstract: The poster shows an annotated but mainly visual map of the world, which highlights the activity of R users from various points of view – similar to what I have presented at the previous useR! conference. The plots and the infographics were created in R, inspired by some recent blogposts of cartograms and the xkcd package, but using a wide variety of date sources collected, cleaned, merged and aggregated by the author. These include the number of visitors of R-bloggers.com, the attendees of all previous useR! and some other R-related conferences, the members and supporters of the R Foundation, the number of users on GitHub with R repositories, package download statistics from CRAN mirrors and the number of R User Groups and the number of attendees of the events. Besides these raw data, the poster will also present a population-weighted scale of R activity for all countries of the world.

Keywords: R activity, R User Groups, CRAN statistics, cartogram, spatial data

R's deliberate role in Earth surface process research

Michael Dietze

Helmholtz Centre Potsdam - German Research Centre for Geosciences

http://www.gfz-potsdam.de

Abstract: Geomorphology, i.e. the investigation of processes that shape our planet at scales of milliseconds to millions of years, of nanometers to continents, is a long-established, vibrant scientific field, which is confronted with transdisciplinary methodological demands.

Two key innovations in the last 30 years have profoundly affected this discipline: the ability to efficiently quantify the shape and rate of change of landforms and ii) the ability to determine the ages of landforms. This has renewed the relevance of Earth surface process research to many disciplines, such as geology, biology, geography, engineering and social sciences. The integrative, scale-bridging nature of geomorphology demands matching data handling methods.

This contribution will give insight to emerging Earth surface process research fields and how R feeds into comprehensive and effective data processing. It shows how existing package functionalities contribute to novel, task-specific packages. The contribution highlights how more and more packages cover methodological gaps and how CRAN policy allows joining these packages to integrated spatial data handling, time series analysis, numerical modelling, signal processing as well as statistic analysis and modelling.

Keywords: Geomorphology, End-member modelling, Landscape evolution modelling, Environmental seismology, Dating

remote: Empirical Orthogonal Teleconnections in R

Florian Detsch

Environmental Informatics, Philipps-Universität Marburg, Germany
http://environmentalinformatics-marburg.de/

Abstract: 'remote' implements a collection of functions to facilitate empirical orthogonal teleconnection analysis. Empirical Orthogonal Teleconnections (EOTs) denote a regression-based approach to decompose spatio-temporal fields into a set of independent orthogonal patterns. They are quite similar to Empirical Orthogonal Functions with EOTs producing less abstract results that are orthogonal in either space or time. In this paper we present the R implementation of the original algorithm by Huug van den Dool in the 'remote' package. Especially the utilisation of Rcpp for the intensive regression calculations ensures acceptable computation times and memory usage for this 'brute force' spatial data mining algorithm. This is a very important aspect of 'remote' as the amount of data points in spatio-temporal geoscientific fields is generally extremely large and can easily require millions, or even billions of calculations. To highlight its usefuleness we provide some examples of potential use-case scenarios for the method including the replication of one of the original examples from van den Dool's original paper, as well as statistical downscaling from coarse to fine spatial grids (using NDVI fields).

Keywords: teleconnection analysis, spatial data mining, raster data, Rcpp, R

Reproducibility in environmental modelling

Michael Rustler, Christoph Sprenger, Nicolas Caradot and Hauke Sonnenberg

Kompetenz Wasser, Berlin

http://www.kompetenz-wasser.de/

Abstract: In environmental sciences numeric models play an important role supporting decision making. Usually, the modelling procedure (parameterisation, calibration, validation, scenario analysis) includes manual steps. For example, models are often calibrated by changing parameter values manually using a trial-and-error method (e.g. Anibas et al. 2009). This makes it challenging to document how the modelling software was used, which is a prerequisite for making the applied methodology transparent and thus the whole modelling process reproducible.

Automatisation by means of programming can improve the modelling process. Once a methodology is implemented in the form of program code it is inherently documented. The code can be run repeatedly and will always produce the same results given the same inputs.

We used R as programming language to automate the modelling workflow.

Different models have been 'wrapped' by means of R packages: VS2DI (groundwater flow, heat and solute transport), WTAQ-2 (well drawdown), EPANET (pressurised pipe networks) and Gompitz (sewer ageing).

These models can now be configured and run from within the R environment. This allows to use R's excellent functions for retrieving and preparing input data (e.g. monitoring, geographical data) as well as analysing and plotting simulation results and generating reports.

Modelling is described in the form of version controlled R scripts so that its methodology becomes transparent and modifications (e.g. error fixing) trackable. This leads to reproducible results which should be the basis for smart decision making.

Keywords: modelling, reproducibility, automatisation

Reproducible Statistics course for the future, from Stata to R

Kennedy Mwai

Data Manager, KEMRI-Wellcome Trust Research Programme

http://kemri-wellcome.org/

Abstract: Reproducibility being laudable and frequently called for, we should be instilling this practice in students before they set out to do research. The maturity and extensive reproducibility abilities of Git, R and RStudio based materials make an excellent choice for professional statistical skills training. The most commonly used statistical softwares namely SAS, SPSS and STATA can cost over US500$for a single license and over US$5,000 for a discounted twenty user access and can be challenging to create reproducible courses for universities and training institutions. Currently, it has been observed that the trend is changing and R popularity is rising. This talk will focus on the success of converting a two weeks statistical methodology for the design and analysis of epidemiological studies (SMDAES) course from Stata based to R based reproducible course using RMarkdown or LaTeX, Git and GitHub, R and RStudio server. We will present the importance of using Git ,R and RStudio as tools for statistical workshops and institution trainings over commercial based tools.The talk will also show the success of teamwork on Git while creating the course materials.

Keywords: Training, Reproducibility, RMarkdown, Git

RJSDMX: accessing SDMX data from R

Attilio Mattiocco

Economics and Statistics Department, Bank of Italy, Rome, Italy

http://www.bancaditalia.it

Abstract: SDMX (Statistical Data and Metadata Exchange) is a standard for the exchange of statistical data. It is widely adopted by international institutions (e.g. ECB, OECD, Eurostat, IMF and others) for data dissemination.

The RJSDMX package has been built as a connector between SDMX data providers and the R environment. The package provides functions for connecting to SDMX web services and downloading data as zoo time series. The package also provides text and graphical functions for exploring the metadata contents of the providers, helping the user identify the data of interest, build and validate specific queries. The package is part of the Web Technologies Task View and it has been inserted in the rOpenSci project.

The RJSDMX package is part of a wider framework, the 'SDMX Connectors for Statistical Software', an Open Source project that aims to provide the same SDMX data exploration and access functions in the most popular data processing tools (R, STATA, SAS, Excel, MATLAB).

Keywords: SDMX, web services, web technologies, data access

saeSim: Simulation Tools for Small Area Estimation

Sebastian Warnholz

Statistische Beratungseinheit

http://www.stat.fu-berlin.de/

Abstract: The demand for reliable regional estimates from sample surveys has substantially grown over the last decades. Small area estimation provides statistical methods to produce reliable predictions when the sample sizes in specific regions are too small to apply direct estimators. Model- and design-based simulations are used to gain insights into the quality of the methods utilized. We present a framework which may help to support the reproducibility of simulation studies in articles and during research. The R-package saeSim is adjusted to provide a simu- lation environment for the special case of small area estimation. The package may allow the prospective researcher during the research process to produce simulation studies with minimal coding effort. It provides a consistent naming convention and highlights a literate programming philosophy.

Keywords: Simulation, Small Area Estimation, saeSim

Semi-Supervised Learning in R

Jesse H. Krijthe

Pattern Recognition and Bioinformatics, Delft University of Technology, The Netherlands

http://prb.tudelft.nl/

Abstract: Semi-supervised learning considers a particular kind of missing data problem, one in which the dependent variable (label) is missing. The goal of semi-supervised learning is to construct models that improve over supervised models that disregard the unlabeled data. These models are used in cases where unlabeled data is easy to obtain or labeling is relatively expensive. Example applications are document and image classification and protein function prediction, where additional objects are often inexpensive to obtain, but labeling them is tedious or expensive.

For my research into robust models for semi-supervised classification I have implemented several new and existing semi-supervised learners that have been combined in the RSSL package. This package also contains several functions to set up benchmarking and simulation studies, comparing several semi-supervised algorithms. This includes the generation of different kinds of learning curves and cross-validation results for semi-supervised and transductive learning. The goal of this work is to make reproducible research into semi-supervised methods easier for researchers and to offer simple consistent interfaces to semi-supervised models for practitioners.

The package is still under development and I would like to discuss how to improve the interfaces of the models to interact more easily with other packages.

Keywords: Semi-supervised learning, Machine Learning, Classification, Missing data

Shiny Application Using Multiple Advanced Techniques

Ann Liu-Ferrara

Statistical Tools Department, BD

http://www.bd.com/

Abstract: Demonstrate a shiny application that uses multiple new technologies to increase efficiency and user ease. The tool was created to combine material testing results from a pdf and up to five data files and process the results for upload into a database. The tool automates tedious manual work that had taken users hours to do, and reduces the chance of human error. The tool is for BD internal use, but the techniques have wide applicability.

Behind the user interface the reactive upload fields are created within a sapply loop which is efficient and reduces replicate code. The data in each file are displayed as the data are loaded using a ggvis plot. Each plot contains either one or five curves showing the testing results, one curve per sample. The user can see data details by hovering the mouse over a data point and can select a curve by clicking. The selected curve is cleaned and will be highlighted instantaneously. The user can download the raw and cleaned data and multiple graphs for upload to the database. This result is in an easy to use, fast interface that produces a dramatic time saving and a very satisfied client.

Keywords: ggvis, shiny, click, loop, automation

Short-term forecasting with factor models

Rytis Bagdziunas

CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

https://www.uclouvain.be/en-core.html

Abstract: Macroeconomic data has lately become accessible in computer-friendly formats, e.g SDMX REST APIs used by Eurostat, ECB or OECD. Such data availability allows analysts and researchers keep and maintain their datasets up to date at no cost and with little effort.

Easy access to large datasets continues to spur growth of data-driven techniques for short-term forecasting and construction of diffusion indices. While these techniques are still being actively researched, common dimension reduction methods, such as principal component or factor analysis, are already known to be consistent and have desirable properties under fairly reasonable assumptions, even for serially correlated economic data. `dynfactoR` package (in development) aims to reimplement these economic models in `R` language in a concise and self-explanatory manner. This includes dynamic factor estimation based on EM algorithm and Kalman filtering, support for missing data, linking quarterly and monthly observations as well as data segmentation.

My poster will illustrate how, in a few minutes, `dynfactoR` along with `rsdmx` package can be used to construct from scratch easily maintainable short-term economic forecasting models loosely comparable to those used nowadays in central banks.

Keywords: factor models, forecasting, nowcasting, factor models, sdmx

Sparkle - Deploying Shiny Apps at AdRoll

Maxim Dorofiyenko

AdRoll

Abstract: Shiny is an amazing tool for building powerful web applications with R but the deployment process can be a real headache. Normally developers need to interact with a command line (ssh/scp) to work on their apps. Sparkle is a framework for deploying shiny apps that promotes version control and attempts to eliminate barriers to shiny development. We do this using a combination Jenkins CI and the R packages Brew and Knitr. With the Sparkle workflow deploying an app becomes as easy as pushing to Github. From there the developer can easily iterate on their app without running into some of the common problems associated with shiny app hosting on a remote machine. The goal is to make this available through open source in the near future.

Keywords: Shiny, Knitr, app development, Brew, Jenkins

Statistical Analysis Problems in Fair Lending Regulation

Bruce Moore

Moore Software Services LLC

https://www.mooresoftwareservices.com

Abstract: Differences in loan interest rates for different racial and ethnic groups in the United States have been a topic of research for decades and are currently used as the basis for regulatory enforcement actions under Fair Lending laws. There is significant public policy debate over the statistical methods used by regulatory agencies and the method that is used to estimate the race and ethnicity of borrowers, as lenders cannot legally require race and ethnicity information on a loan application.

A simulation model shows that the lower accuracy of race and ethnicity estimation for African Americans makes it unlikely that race-based discrimination would be detected. It also shows that it is likely that a small number of enforcement actions would occur when perfect race and ethnicity identification would detect no race-based discrimination.

A second simulation model shows that a very small number of preferential loans to white non-Hispanics can result in a an enforcement action when the total number of loans to white non-Hispanics is small relative to other racial and ethnic groups.

Keywords: simulation, ethnicity, finance, discrimination, policy

Statistical Approaches to Corpora Analysis (NLP)

Patrick Bolbrinker

no affiliation (private)

http://no.affiliation.private.org

Abstract: Objective: Present a computational workflow in R for text exploration and give an overview of possibilities and problems pertaining to quantitative text analysis.

Background: Natural language processing (NLP) started with word length studies in the late 19th century. Today, over 100 years later, corpora analysis is - besides the development of efficient machine translation systems - one of the main objectives in natural language processing. Context processing and analysis of texts derived from large sources, e.g. all tweets in 2014, can be extremely difficult or even impossible due to limited computational recourses. Therefore, it is more reasonable to use a representative subset. However, how do we measure "appropriate" for sample selection, and deal with language ambiguity?

Methods: With the novice in mind I present an efficient workflow for preprocessing, sampling and statistical analysis of English language texts from Twitter and Project Gutenberg. The main focus lies on statistics for randomized sample selection and quantitative corpora exploration (discourse analysis, text-to-text word "marker" distributions etc.).

Keywords: NLP, corpus exploration, descriptive statistics

sValues: a package for model ambiguity in R

Carlos Cinelli

Central Bank of Brazil and University of Brasilia

http://www.bcb.gov.br/pt-br/paginas/default.aspx

Abstract: It is common to have many potential explanatory variables to choose from in situations in which the theory can be ambiguous about which ones to include in the model. One way to tackle this problem is using Bayesian Model Averaging, and the R ecosystem has (among others) two good packages for that, BMA and BMS. This presentation will introduce the sValues package (soon to be on CRAN), which provides an implementation of the S-value statistic, a measure of sturdiness of regression coefficients proposed by Leamer (2014a) and discussed in Leamer (2014b) to assess model ambiguity, an alternative approach to the methods above mentioned. The sValues package has a main function (with formula, data.frame and matrix methods) which does all the analysis and calculations for the user, and it also provides methods for summary, coefficients, plots and printing to let the user explore and export the results. To illustrate the package use, we use the "Growth Regressions" example, showing how one can easily replicate Leamer (2014a) using the sValues package and also compare its results with those of Fernández et al (2001, Sala-i-Martin et al (2004) and Ley, E. and Steel, M. F. (2009) using the BMS package.

Keywords: Model Uncertainty, Bayesian Statistics, Bayesian Model Averaging, S-Values, Extreme Bounds

The 7 quality control tools in a nutshell: R & ISO approaches

Emilio L. Cano

Universidad Rey Juan Carlos

http://www.urjc.es

Abstract: Quality Control is a Statistics application field that can be useful to any activity sector: not only manufacturing, but also services and administration. As a mainly business-oriented methodology, commercial software has been prominent for the application of statistical quality control techniques. However, the generalization of R as statistical software within companies is also reaching this field. The power of R allows to perform any task in a commercial software, and even more through both the base and contributed packages.

This presentation is part of a forthcoming book in Springer's Use R! series and shows the seven basic quality control tools and how to apply them using R. They were named by Kaoro Ishikawa after the seven weapons used by the warrior monk Benkey, who was able to succeed in battles with just those seven weapons. Similarly, Ishikawa stated that 95% of the problems at the shop-floor level can be solved with the seven basic quality tools, if they are used properly. The seven tools are: cause-and-effect diagrams, checksheets, control charts (the rock star), histograms, Pareto charts, scatter plots, and stratification.

The relationship of the tools with ISO Standards complete the contents of this work.

Keywords: Quality Control, Six Sigma, ISO Standards, Applied Statistics, Ishikawa

The biogas package: simplifying and standardizing analysis of biogas data

Sasha D. Hafner, Charlotte Rennuit

Department of Chemical Engineering, Biotechnology and Environmental Technology, University of Southern Denmark, Odense, Denmark

http://www.sdu.dk/en/om_sdu/institutter_centre/ikbm_kemi_bio-_og_mijoeteknologi

Abstract: Anaerobic digestion is a biological-based process for producing biogas (a mixture of methane and carbon dioxide) from organic material. It is an important source of renewable energy. For example, Denmark has > 20 centralized biogas plants converting organic waste into heat and electricity. Tens of millions of small digesters produce cooking and lighting fuel from household waste in China. Biogas production is an active research area, and laboratory experiments are used to quantify production from particular substrates or systems. Collected data must be processed in a set of steps which may be implemented in different ways. These steps are rarely fully described, complicating comparisons between experiments. We developed an R package, "biogas" (available from CRAN), to simplify data analysis and increase reproducibility. Low-level functions include stdVol for standardizing gas volumes and interp for interpolating biogas composition or production. The cumBg function can be used to calculate cumulative gas production and rate, combining interpolation, volume standardization, and summation. Biochemical methane potential (BMP) can be directly obtained using the flexible summBg function. And biogas production and composition can be predicted using predBg. We hope that the biogas package simplifies analysis of biogas data and facilitates standardization in data processing.

Keywords: Data manipulation, Biogas, Laboratory data

The conduit Package

Ashley Noel Hinton and Paul Murrell

Department of Statistics, The University of Auckland, Auckland, New Zealand
https://www.stat.auckland.ac.nz/

Abstract: The 'conduit' package for R is intended to support greater use of open data sets by encouraging the creation, reuse, and recombination of small scripts that perform simple tasks. The package provides a "glue system" for running "pipelines" of R scripts. Each script is embedded in an XML "module" wrapper, which defines the inputs required by the script and the outputs that the script produces. An R script can be made into a module even when the original script author has no knowledge of 'conduit'. This means that it is easy for any script to be reused in pipelines via the 'conduit' package. Embedding scripts in modules also helps to organise code and makes it simple to reuse scripts in other pipelines. Pipelines specify connections ("pipes") from the outputs of one module to the inputs of another module. The 'conduit' package orchestrates the execution of the module scripts and passes their results to subsequent modules. As modules are defined in XML, they are not specific to R scripts, and it is possible to wrap scripts written in other languages, such as Python.

Keywords: open data, data pipelines, code reuse, code organisation, XML

The dendextend R package for manipulation, visualization and comparison of dendograms

Tal Galili

Tel Aviv University

http://www.math.tau.ac.il/index.php?Itemid=27

Abstract: A dendrogram is a tree diagram which is often used to visualize a hierarchical clustering of items. Dendrograms are used in many disciplines, ranging from Phylogenetic Trees in computational biology to Lexomic Trees in text analysis. Hierarchical clustering in R is commonly performed using the hclust function. When a more sophisticated visualization is desired, the hclust object is often coerced into a dendrogram object, which in turn is modified and plotted. The dendextend R package extends the palette of base R functions for the dendrogram class, offering easier manipulation of a dendrogram's shape, color and content through functions such as rotate, prune, color_labels, color_branches, cutree, and more. These can be plotted in based R and ggplot2. dendextend also provides the tools for comparing the similarity of two dendrograms to one another: either graphically (using a tanglegram plot, or Bk plots), or statistically (with Cophenetic correlation, Baker's Gamma, etc) - while enabling bootstrap and permutation tests for comparing the trees. The dendextendRcpp package provides C++ faster implementations for some of the more computationally intensive functions.

Keywords: dendrogram, visualization, clustering, hierarchical clustering, ggplot2

The Preludes to Civil War: Analyzing the Factors in R

Jefferson Davis

Research Analytics, Indiana University, Bloomington, Indiana, USA
http://rt.uits.iu.edu/

Abstract: The factors that encourage violence and civil war are of clear interest. Work in the last decade has suggested that the most important are poverty, political instability, rough terrain, and large populations. This downplays ethnic divisions as well both the role democracy plays in providing a non-violent outlet for dissidence and the role that authoritarianism plays in tamping down dissidence.

Using R to access national datasets can clarify the source of the data and allow easy updating. R scripts also automate the standardization of names for nations, ethnic groups, and religious groups within nations. A scripting approach also clarifies judgment calls in political science: should a state and a successor state count as separate nations? At what point does a breakaway region establish independence? What level of conflict qualifies as a war? All these questions allow multiple sensible answers. It is best to make the choices clear and easy to change if warranted.

In our analysis both stable democracy and authoritarianism decrease the risk of civil war, refining the role of political instability as a factor. Also, with current detailed geographical data the rough terrain effect disappears. The use of R makes this analysis more transparent and more widely replicable.

Keywords: Intrastate war, measuring conflict, fractionalization, measures of democracy, national level datasets

The seqHMM package: Hidden Markov Models for Life Sequences

Satu Helske

Department of Mathematics and Statistics, University of Jyväskylä, Finland
http://www.jyu.fi/

Abstract: In social sciences, sequence analysis is being more and more widely used for the analysis of longitudinal data such as life courses. Life courses are described as sequences, categorical time series, which constitute of one or multiple parallel life domains. Sequence analysis is used for computing the (dis)similarities of sequences and often the goal is to find patterns in histories using cluster analysis. However, describing, visualizing, and comparing large sequence data with multiple life domains is complex. Hidden Markov models (HMMs) can be used to compress and visualize information by detecting underlying life stages and finding clusters.

The seqHMM package is designed for the HMM analysis of life sequences and other categorical time series. The package supports models for one or multiple sequences with one or multiple channels (dimensions/life domains), as well as functions for model evaluation and comparison. Sequence data can be clustered during the model fitting and external covariates can be added to explain cluster membership. Visualization of data and models has been made as convenient as possible. The user can easily plot multichannel sequence data, convert multichannel sequence data to single channel representation, and visualize hidden Markov models.

Keywords: life sequences, categorical time series, hidden Markov models, data visualization, model visualization

The VALOR package: Vectorization of AppLy for Overhead Reduction of R

Haichuan Wang

Computer Science Department, University of Illinois at Urbana-Champaign
http://cs.illinois.edu/

Abstract: In this presentation, we will introduce the VALOR package. Its purpose is to improve the performance of program written in terms of Apply. The implementation of Apply in the R interpreter incurs in significant overhead resulting from the iterative application of the input function to each element of the input data. For example, in lapply(L,f), the function f will be interpreted once for each element of L.

Our approach performs data transformation and function vectorization to convert the looping-over-data execution into vector operations. The package transforms the input data to Apply operations into vector form and vectorizes the function invoked. In this way, it converts the conventional, iterative implementation of Apply into a single function invocation applied to a vector parameter containing all the elements of the input list. With the built-in support of vector data types and vector operations, this new form has much less interpretation overhead since the vectorized function is only interpreted once and vector operations in it are supported by native implementations.

We used a suite of data analysis algorithm benchmarks to evaluate the package. The results show that the transformed code can achieve on average 15x speedup for iterative algorithms and 5x for direct (single-pass) algorithms.

Keywords: program optimization, lapply, performance, interpreter overhead, vectorization

Type-I error rates for multi-armed bandits

Markus Loecher

Dept. of Economics, Berlin School of Economics and Law, Germany

http://www.hwr-berlin.de/en/department-of-business-and-economics/academic-staff/vcard-single-en/?tx_wmdbvcard_pi1[show]=2184&cHash=c56c754d093f312721bc8bbc7eec617a

Abstract: The name "multi-armed bandit" describes a hypothetical experiment where one has to choose between several games of chance ("one-armed bandits") with potentially different expected payouts. This is a classic "exploit-explore" dilemma as one needs to both find the game with the best payout rate, but at the same time maximize one's winnings. It was recently claimed that in the context of ad campaign optimization sequentially updating Bayesian posterior probabilities in combination with Thompson sampling would enable decision making at drastically reduced sample sizes compared to a classical hypothesis test. We have implemented these ideas in the R package bandit.

We further derive a normal approximation to the quantile of the value-remaining distribution which is fast to compute. The agreement is within a few (relative) percent. We further show empirical results which indicate that the stopping rule based on the value-remaining distribution is overly optimistic. The true rate of falsely declaring the wrong arm to be the winner is substantially higher than the set level would suggest. We speculate that this "significance level" needs to be adjusted due to multiple testing.

Keywords: multi-armed bandit, sequential learning, multiple testing, type-I error

Using machine learning tools in R to project the frequency of high-mortality heat waves in the United States under different climate, population, and adaptation scenarios

Brooke Anderson

Department of Environmental & Radiological Health Sciences, Colorado State University, Fort Collins, Colorado, USA

http://colostate.edu

Abstract: Certain rare heat waves can have devastating effects to a community's public health and well-being. Here, we used R machine learning tools to build models that classify a heat wave as "very dangerous" to human health or "less dangerous" based on heat wave characteristics (e.g., absolute temperature, temperature relative to its community's temperature distribution, length, timing in season). Very dangerous heat waves are very rare, and so we considered methods to account for this class imbalance in model building. To build and test these models, we used data from 82 large US communities, 1987—2005. We built and evaluated five different types of models (two types of classification trees, bagging, boosting, and random forests ensemble models) using four different approaches for class imbalance (nothing, over-sampling from the rare class, over / under-sampling, and Randomly Over Sampling Examples [ROSE]), for a total of twenty models. We evaluated the models with Monte Carlo cross-validation, identifying three acceptable models. Using these, we predicted the frequency of very dangerous heat waves in these 82 communities in 2061—2080 under two scenarios of climate change (RCP4.5, RCP8.5), two scenarios of population change (SSP3, SSP5), and three scenarios of community adaptation to heat (none, lagged, on-pace).

Keywords: climate change, machine learning, classification models, epidemiology, heat waves

Using R for allergy risk assessment in food product

Sophie Birot

Statistics and Data Analysis Section, DTU Compute, Danish Technical University, Denmark

http://www.compute.dtu.dk/english/research/Stat

Abstract: Food allergies are a public health concern as high prevalence and severity of the reaction can lead to harmful consequences. So, risk management must be conducted; avoidance diets are the most common way for allergen management. However, a risk might remain due to allergens contamination of food products leading to an unintended consumption of allergen. To estimate the risk following unintended allergen consumption, the recommended approach is the probabilistic risk assessment. It is currently reviewed and improved within the iFAAM project (Integrated Approaches to Food Allergen and Allergy Risk Management). This method takes into account 3 different sources of information: the amount of unintended allergen in the food (product contamination), the consumption of the contaminated product and the allergen threshold distribution (allergen dose which triggers an allergic reaction). All 3 distributions are modelled using the R software. Risk simulations are performed 2 different ways in R; Monte Carlo simulations and Bayesian networks to assess the number of allergic reaction. These methods both propagate variability and uncertainty from the input variables to the outcome, hence confidence and credibility intervals are provided for the allergic risk.

Keywords: food allergy, risk assessment, Monte Carlo simulations, Bayesian analysis, uncertaintyand variability propagation

Using R to analyze how R is being used

Stanislaw Swierc

Institute of Computer Science, Silesian University of Technology, Poland
http://www.polsl.pl/en/Strony/AutomaticControl.aspx

Abstract: R is a very popular language of choice among statisticians and data scientists who share their work with the Open Source Software community. They do it by making their code available either by publishing it for immediate download or by hosting it in public repositories. GitHub is one of the biggest project hosting service where many R packages such as devtools are being developed. This platform is open and it has been used as a source of data for research on OSS development.

In our research we mine GitHub for projects that use R. There are over 75k public repositories with list it as their primary programming language. They make heavy use of the self-contained packages available from The Comprehensive R Archive Network. By looking at which packages are referenced together in existing code we can extract frequent itemsets and association rules to discover interesting relationships and usage patterns. The end-to-end research, which includes data collection, analysis and visualization, is performed using many tools available to R users.

Keywords: arules, association rules, visualization, github

Using R to build a coherence measure between LISA functions and its use for classification in spatial point patterns

Francisco Javier Rodríguez Cortés

Department of Mathematics, Jaume I University, Castellón, Spain

http://www.uji.es/CA/departaments/mat/estructura/personal/e@/22752?p_per_id=338537

Abstract: Modelling real problems through spatial point processes becomes essential in many scientific fields. Spatial cluster analysis is a key aspect of the practical analysis of spatial point patterns. The idea of considering individual contributions of a global estimator as a measure of clustering was introduced by Anselin (1995) with the name of Local Indicators of Spatial Association (LISA), and it has been used as an exploratory data analytic tool to examine individual points in a point pattern in terms of how they relate to their neighbouring points.

The local versions of the second-order product density set a powerful tool to address the problem of classification of interesting subpatterns that often form spatial clusters. LISA functions can then be grouped into bundles of similar functions using multivariate hierarchical clustering techniques according to a particular statistical distance Cressie and Collins (2001).

We introduce a new coherence measure for the classification of bundles of LISA functions to classify points according to a certain clustering degree in the pattern. The performance of this technique is outlined through multivariate hierarchical clustering methods and multidimensional scaling using R. We apply this methodology for the classification of Earthquake Catalog on a seismically active area.

Keywords: Coherence measure, Fourier transform, Local indicator of spatial association, Second-order product density

Using R to improve compliance in clinical trials

Luke Fostvedt

Pfizer Inc.

http://www.pfizer.com

Abstract: Clinical trials have many moving parts and protocol compliance is important to gain the necessary knowledge for a comprehensive submission package to regulatory agencies. The collection of both pharmacokinetic (PK) and pharmacodynamics (PD) samples is the basis for characterizing the safety and efficacy of any new compound. In some cases, there can be a misunderstanding between the pharmacologists and the clinicians as to why the timing of specific collections is important. We have seen that graphics connecting the sampling schedule with PK time/conc curves improves the understanding of why such collections are necessary. R provides a very flexible platform to share this information with clinicians. Using R and shiny clinicians are able to understand the expected behavior of the drug adsorption and disposition. The physicians are also able explore the behavior under many different assumptions (amount dosed, frequency of administration, linear vs. nonlinear elimination, two-compartment v. multiple compartment models, etc.) and patient characteristics (weight, height, age, body-surface area, renal and hepatic impairment, etc.) that could be important factors to consider for dose adjustments. R graphics along with a shiny app will be presented to illustrate how compliance can be improved.

Keywords: Clinical Trials, shiny, graphics, FDA, industry applications

Value-added indicatoRs for schools: using R for school evaluation in Poland.

Tomasz Żółtak

Educational Research Institute, Warsaw, Poland

http://www.ibe.edu.pl/en/

Abstract: Educational value-added measures attempt to evaluate school and/or teacher quality. They are used in various forms and on a large scale mostly in the US (e.g., TVAAS/EVAAS) and the UK (as an element of School League Tables). A few years ago they were implemented also in Poland, providing indicators for about six thousand lower-secondary schools and five thousand upper-secondary schools a year (see http://ewd.edu.pl/en/). We would like to present how R has been integrated into a heterogeneous system computing value-added indicators for Polish schools and making the indicators available to the public. Computing value-added indicators is a complex process which involves scaling examination scores with IRT models, estimating mixed-effects regression models for large data sets and sophisticated data manipulation. We use R in three ways. First, it is our primary statistical tool (especially the packages mirt and lme4). Second, it operates external software (Mplus). Third, it integrates the analytical process with a huge SQL database, i.e. it retrieves data from the database, conducts statistical analyses and stores their results in the database. In sum, we would like to share our experiences in applying R to complex solutions in public information systems.

Keywords: value-added indicators, R as a glue language, public indicators

vdmR : Web-based visual data mining tools by multiple linked views

Tomokazu Fujino

Department of Environmental Science, Fukuoka Women's University, Japan
http://www.fwu.ac.jp/

Abstract: The vdmR package generates web-based visual data mining tools by adding interactive functions to ggplot2 graphics. Brushing and linking between multiple plots is one of the main features of this package. These functions are well known as "multiple linked view". Currently, scatter plots, histograms, parallel coordinate plots, and choropleth maps are supported in the vdmR package. In addition, identification on the plot is supported by linking between the plot and the data table. In this talk, we will introduce the basic usage of this package and give some demonstrations of implementing this package as a Web application.

Keywords: data visualization, multiple linked views, interactive graphics, ggplot2, choropleth maps

Web Structure Mining Using R

Roy Smith

College of Science & Health Professions, Northeastern State University, Tahlequah, Oklahoma, United States

http://academics.nsuok.edu/sciencehealth/ScienceHealthHome.aspx

Abstract: As both consumer and personal websites expand in complexity, the structure of these sites can become convoluted. As such, human-computer interaction becomes extremely important for usability. This project uses a combination of technologies to implement a Web structure mining procedure to generate raw data for a given website. An analyzer, written in R, has been created to assimilate and generate a 3D visual "tree" of the site. This tree will allow for users to have an easily understandable sitemap. An additional feature allows these maps to be exported to a format supported by a 3D printer to enable users to create a physical model. Using additional raw data, links can be converted into occurrence listings and used to generate plots that allow simple comparisons of each link to every other link from the given site. It is the intention that these sitemaps and occurrence plots can be used in comparison with maps from other sites for designers to easily determine how to restructure a site to have a more efficient layout for users. With further time and research, this system may also be used to find patterns across the internet to determine the separation of important links.

Keywords: R-analyzer, web structure mining, sitemap, human-computer interaction, 3D representation

Web-scraping with R - Collecting Data from Facebook

András Tajti

Department of Statistics, Eötvös Loránd University, Hungary
http://statisztika.tatk.elte.hu/

Abstract: Facebook has over one billion users, events, pages, etc., with a lot of personal information about them, which makes the site the worlds largest information repository. The company provides an API to connect to its services and informations to retrieve interesting data about its members. Although there is an R package for communicating this API (RFacebook), the interface itself became so restricted that it can provide less information for experienced users, and almost nothing to a layman, compared to what anyone can see through logging in as a user. R has a package for simulating browser usage (RSelenium) based on the Selenium software testing framework, which enables every user to write R scripts for web surfing. Using this technology, I created functions which can be used to browse Facebook - especially search, gather profile info or collect friendships - through R for the non-expert R users without any knowledge of javascript or the Selenium Web Driver. Also, this way from an R session one can see Facebook from the logged-in user's eye view, and not through an API's small scope. These functions are only experiments as this kind of data collection is governed by the Automated Data Collection Terms.

Keywords: web technologies, web-scraping, RSelenium, Facebook, automation

Who is afraid of R? Strategies for overcoming faculty resistance in using R in business curriculums

Gokul Bhandari

Management Science, University of Windsor, Ontario, Canada

http://www1.uwindsor.ca/odette/gokul-bhandari

Abstract: The purpose of this presentation is to demonstrate how R can be successfully introduced in business curriculums by implementing various strategies. A shinyapps developed for analyzing Assurance of Learning (AOL) will be demonstrated and discussed.

Keywords: Lewin's Change Management Theory, Curriculum design and pedagogy, Assurance of Learning, Shinyapps

Part V

Sponsor Session

Sponsor Session

DataRobot: DataRobot API

Ted Kwartler

DataRobot

http://www.datarobot.com

Abstract: In this presentation we will present the power of the DataRobot API, a programmatic access to a massively parallel modeling engine. The presentation will include use cases and examples.

RStudio: What's New at RStudio? A Snapshot of our Latest Products and Packages

Tareef Kawaf

RStudio

http://www.rstudio.com

Abstract: Since the last useR! conference at UCLA, RStudio has been making investments in a variety of areas that make it easier for data scientists and analysts to create and share their analyses with the world. In today's presentation we will give you a flavor for the work we have done and leave you with pointers so you can explore the items that are of deeper interest. We will provide a high level view of the various investments in the IDE, new R packages, enhancements to existing packages, and the newest options for enterprises deploying R and Shiny in production.

Teradata: Scaling R for Big Data

Venkatesh Sellapa

Teradata

http://www.teradata.com

Abstract: The world of R has collided with big data introducing the challenge: How can organizations analyze massive volumes of data working within memory limitations? In this session, you will learn how Teradata addresses this challenge with high speed, scalable R solutions that allows Global 1000 companies to tackle big data business problems. Do you want to work with innovative solutions? Join the Teradata team.

Revolution Analytics: R at Microsoft

David Smith

Revolution Analytics

http://www.revolutionanalytics.com

Abstract: In April this year, Revolution Analytics became a Microsoft company. In the announcement, Microsoft said it would "build R and Revolution's technology into our data platform products so companies, developers and data scientists can use it across on-premises, hybrid cloud and Azure public cloud environments". In this short talk I will share some progress that has been made at Microsoft on integrating R, and provide some details on what you can expect in the future.

alteryx: Who We Are, What We Do, What We Do for R

Dan Putler

alteryx

http://www.alteryx.com

Abstract: Organizations have an increasing amount of data that can be converted into information that offers the ability to make better decisions, find new opportunities, and improve efficiency. R is a critical advanced analytics tool that many organizations use to turn data into usable information. While the R language is comparatively easy to learn, it is still a traditional, written, computer programming language. Unfortunately, this fact alone limits the potential diffusion of R across most organizations, reducing its potential benefit to an organization. Alteryx is a platform for data blending and advanced analytics. Its objective is to empower a greater number of individuals across an organization to successfully accomplish these tasks, improving the overall performance of an organization. Alteryx uses R as a key element in powering much of its advanced analytics capabilities, in a much easier to approach interface. Alteryx itself is best viewed as both a data pipelining engine and a visual programming framework. In this talk, we introduce ourselves, highlight the benefits we provide our customers, particularly as it relates to R, and cover some of the things we do to help support the R community.

TIBCO Spotfire: Extending the Reach of the R Language to the Enterprise

Lou Bajuk-Yorgan

TIBCO Spotfire

http://spotfire.tibco.com

Abstract: R provides tremendous value to statisticians and data scientists; however, these users of R are often challenged to extend that value to the rest of their organization. TIBCO, through it's enterprise-class, alternative R interpreter, TIBCO Enterprise Runtime for R (TERR), helps R users share their analytics more widely. TERR draws upon our long history of developing S+, and is integrated into BI applications (through Spotfire), real-time environments (through TIBCO Streambase), and into 3rd party products (such as Lavastorm Analytics), helping R users serve a much wider audience within their organizations. TERR provides an embeddable, high-performance platform for R language analyses.

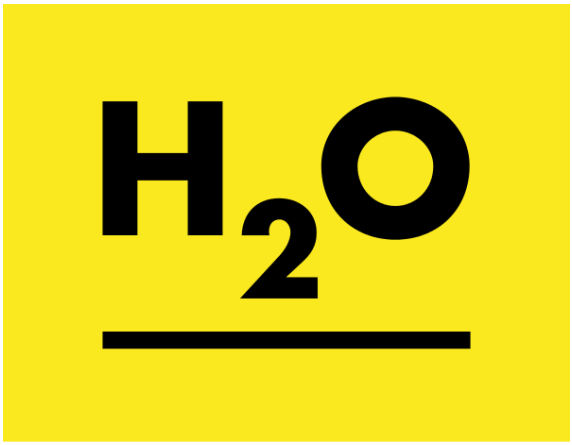
H_2O: Intro to h2o

Amy Wang

H_2O

http://www.h2o.ai

Abstract: H2O is a fast, scalable, open-source machine learning platform for building smarter applications. Customers like PayPal, Nielsen, Cisco and others choose H2O for accurate prediction scenarios and combinations of high volume data with multiple models. H2O's speed enables more iterations from a broad selection of algorithms, including GLM, Random Forest, GBM, and Deep Learning. H2O's easy-to-use APIs allow users to immediately integrate models into R, Python, Spark, Excel or Tableau. The company's customers have built powerful predictive engines for Recommendations, Customer Churn, Propensity to Buy, Dynamic Pricing and Fraud Detection for sectors including Insurance, Healthcare, Telecommunications, AdTech, Retail and Finance



HP: HP Haven Predictive Analytics powered by Open Source Distributed R

Indrajit Roy

HP

http://www.hp.com

Abstract: HP Distributed R is an open-source framework for large-scale machine learning, statistical analysis, and graph processing. It splits tasks among multiple nodes and supports your favorite statistical packages. We will discuss HP's open source effort to make R scale and how you can help add more scalable algorithms to R. Come be part of it!