

THE 8TH INTERNATIONAL *useR!* CONFERENCE

ABSTRACT BOOKLET



VANDERBILT  UNIVERSITY

DEPARTMENT OF BIostatISTICS

ABSTRACT BOOKLET

The abstracts contained in this document were reviewed and accepted by the useR! 2012 program committee for presentation at the conference. The abstracts appear in the order that the presentations were given, beginning with tutorial abstracts, followed by invited and contributed abstracts. The index at the end of this document may be used to navigate by presenting author name.

Reproducible Research with *R*, \LaTeX , & Sweave

Theresa A Scott, MS^{1*}, Frank E Harrell, Jr, PhD^{1*}

1. Vanderbilt University School of Medicine, Department of Biostatistics

*Contact author: theresa.scott@vanderbilt.edu

Keywords: reproducible research, literate programming, statistical reports

During this half-day tutorial, we will first introduce the concept and importance of reproducible research. We will then cover how we can use *R*, \LaTeX , and Sweave to automatically generate statistical reports to ensure reproducible research. Each software component will be introduced: *R*, the free interactive programming language and environment used to perform the desired statistical analysis (including the generation of graphics); \LaTeX , the typesetting system used to produce the written portion of the statistical report; and Sweave, the flexible framework used to embed the *R* code into a LaTeX document, to compile the *R* code, and to insert the desired output into the generated statistical report. The steps to generate a reproducible statistical report from scratch using the three software components will then be presented using a detailed example. The ability to regenerate the report when the data or analysis changes and to automatically update the output will also be demonstrated. We will show more advanced usages and customizations of Sweave for making more beautiful reports and for making research more reproducible, as well as discuss differences between Sweave and knitr. In addition, the tutorial will provide a hands-on session during which a practice report will be generated and modified, useful tips and needed resources/references.

Writing efficient and parallel code in R

Uwe Ligges

TU Dortmund University, Germany, and
Clausthal University of Technology, Germany

Uwe.Ligges@R-project.org

Keywords: efficiency, parallelization

After first steps in any programming language, people typically feel need for more advanced programming skills. The word “efficiency” comes to mind. It is typically recognized in the sense of extremely fast code, but it may also mean (at least for me) that code is readable, reusable, and quickly written and documented. Therefore, the most important meaning of efficiency should be time-efficiency of the programmers work.

Nevertheless, when writing new code, we can try to apply some rules right away that help to avoid extremely slow code. During this tutorial, such basic rules will be shown as well as mechanisms that help to speed up the code. If speed up cannot be solved by code optimization any more, “parallelization” may help. We will discuss how to write code that executes in parallel on several cpu cores or even machines using the new *R* package **parallel**.

Contents:

- introduction
- functions
- efficiency and code profiling
- debugging
- parallel code
- *R* on multicore machines or a cluster

Tutorial: Introduction to Rcpp

Dirk Eddelbuettel^{1,*}, Romain François^{2,*}

1. Debian Project

2. R Enthusiasts

*Contact both authors: RomainAndDirk@r-enthusiasts.com

Keywords: R, C++, Programming, Interfaces

Topic: This tutorial will provide a hands-on and practical introduction to the **Rcpp** package (and other related packages). The focus will be on simple and straightforward applications of **Rcpp** in order to extend R, or accelerate execution of simple functions.

The tutorial will also introduce the **inline** package which permits embedding of self-contained C, C++ or Fortran code in R scripts. We will also cover usage of standard **Rcpp** extension packages such as **RcppArmadillo** for linear algebra, **RcppGSL** for interfacing the GNU GSL via its vector and matrix types, and more.

Prerequisites: Knowledge of R as well as general programming knowledge; prior C or C++ knowledge is helpful but not required.

Equipment: Users should bring a laptop set up so that R packages can be built. That means on Windows, the **Rtools** bundle needs to be present and working, on OS X the **Xcode** package should be installed, and on Linux things generally just work. We can provide limited assistance in getting the required tools installed but the focus of the tutorial on how to use them.

References

- [1] Eddelbuettel, D. and R. François (2010). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40(8), 1–18.
- [2] Eddelbuettel, D. and R. François (2011). *Rcpp: Seamless R and C++ Integration*. R package version 0.9.9.
- [3] François, R. and D. Eddelbuettel (2011). *RcppGSL: Rcpp integration for GNU GSL vectors and matrices*. R package version 0.1.1.
- [4] François, R., D. Eddelbuettel, and D. Bates (2011). *RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library*. R package version 0.2.34.
- [5] Sklyar, O., D. Murdoch, M. Smith, D. Eddelbuettel, and R. François (2010). *inline: Inline C, C++, Fortran function calls from R*. R package version 0.3.8.

Fitting and evaluating mixed models using lme4

Douglas Bates

Department of Statistics, University of Wisconsin - Madison bates@stat.wisc.edu

Keywords: linear mixed models, generalized linear mixed models, nonlinear mixed models, profiling

Mixed-effects models or, more simply, mixed models are statistical models that incorporate both fixed-effects parameters, which apply to an entire population or to well-defined subsets of a population, and random effects, which apply to specific experimental or observational units in the study. The tutorial will introduce mixed-effects models and the **lme4** package for fitting, analyzing and displaying linear mixed-effects models, generalized linear mixed models and nonlinear mixed models with scalar or vector-valued random effects in nested, crossed or partially crossed configurations. We will use recently developed capabilities in lme4 that allow for hypothesis testing on and interval estimation of the model parameters using profiled likelihood.

References

- [1] Bates, D., M. Mächler, and B. Bolker (2011). **lme4**: Linear mixed-effects models using s4 classes. <http://cran.R-project.org/package=lme4>. R package version 0.999375-42.

RHIPE: R and Hadoop Integrated Programming Environment; Tutorial

Jeremiah Rounds*

Purdue University

*Contact author: jrounds@stat.purdue.edu

Keywords: Rhipe, Hadoop, R, MapReduce, Divide & Recombine

Use *R* to do intricate analysis of large data sets via Hadoop. Large complex data sets that can fill up several large hard drives (or more) are becoming commonplace, and many *R* using data analyst will have to confront that reality in the coming years. Data parallel distributed computing paradigms such as MapReduce have emerged as able tools to deal with large data, but until now very little has been done to put those paradigms into the hands of *R* users. **Rhipe** is a software package that allows the *R* user to create MapReduce jobs that work entirely within the *R* environment using *R* expressions. This integration with *R* is a trans-formative change to MapReduce; it allows an analyst to quickly specify Maps and Reduces using the full power, flexibility, and expressiveness of the *R* interpreted language.

In this half-day tutorial, the audience will be introduced to distributed computing, Hadoop, MapReduce, **Rhipe**, and common statistical paradigms that can appear in data parallel algorithms. No prior experience will be assumed, but the ideal participant has an interest in distributed computing, Hadoop, MapReduce, and *R*. For those interested in following along with hands on material, a virtual machine with Hadoop, *R* and **Rhipe** preinstalled will be available for download. More information about **Rhipe** is available at www.rhipe.org.

With data analysis examples, we will introduce Divide and Recombine (D&R) for the analysis of large complex data. In D&R data are divided into subsets in one or more ways. Numeric and visualization methods are applied to each of the subsets separately. Then the results of each method are recombined across subsets. By introducing and exploiting parallelization of data, D&R using **Rhipe** succeeds in making it possible to apply to large complex data almost any existing analysis method already available in *R*.

A Tutorial on building *R* Web Applications with Rook

Jeffrey Horner

Vanderbilt University, Department of Biostatistics
jeffrey.horner@vanderbilt.edu

Keywords: brew, Rook, rApache

Rook is an R package that does three things:

- It provides a way to run R web applications on your desktop with the new internal R web server named Rhttpd.
- It provides a set of reference classes you can use to write your R web applications.
- It provides a specification for writing your R web applications to work on any web server that supports the Rook specification.

This tutorial will cover the basics of web programming in the context of R with Rook. Topics covered will include an overview of the Hypertext Transport Protocol, HTML, and Javascript. R Reference classes will be explained with example code using Rook's convenience classes.

Other topics covered will include strategies for sending and retrieving data through an R web application, utilizing templating systems like brew or Sweave, generating on-demand graphics, construction of web services, etc.

Tutorial: Medical Image Analysis in *R*

Brandon Whitcher^{1,2,*}, Jörg Polzehl³, Karsten Tabelow³

1. Mango Solutions, Chippenham, United Kingdom

2. Imperial College London, London, United Kingdom

3. Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

*Contact author: bwhitcher@mango-solutions.com

Keywords: Compartmental models, data management, DICOM, magnetic resonance imaging, NIfTI, nonlinear regression, visualization

Abstract

Quantitative medical image analysis requires a combination of mathematical and statistical methodology to be applied to industry-defined data formats. Whether we are talking about structural assessment of tissue using computed tomography (CT), metabolic activity as measured using ¹⁸F-FDG in positron emission tomography (PET) or structural and functional measures of the brain in magnetic resonance imaging (MRI). To meet this need there is a growing collection of open-source software solutions for all aspects of data management, image processing, analysis and visualization. This tutorial will introduce packages from the CRAN Medical Imaging task view [4, 5] and apply them to structural and functional MRI data. Each section will provide a step-by-step introduction using imaging data that are available from the public domain.

By the end of the tutorial attendees will be able to:

- Read and write medical imaging data in standard formats.
- Manipulate and visualize medical imaging data.
- Apply summary statistics and statistical models to medical imaging data.
- Know where to find resources for the analysis of medical imaging data in the *R* community.

Outline

- Data import/export using **oro.dicom** and **oro.nifti** [7].
- Analysis of functional MRI using **fmri** [3].
- Analysis of diffusion-weighted MRI using **dti** [1].
- Analysis of dynamic contrast-enhanced MRI using **dcmriS4** [6].

Justification

Opportunities for statistics exist in medical image analysis, specifically MRI, because statisticians have played a limited role to date and there is a distinct lack of public-domain software in the field of medical image analysis. A notable exception is in the field of functional MRI, where brain activity is inferred from changes in the magnetic properties of cerebral blood flow. In order to take full advantage of these opportunities and expand the development of statistical methodology in medical imaging analysis we believe that a two-pronged approach is needed.

1. Introduce the field of medical imaging to statistical practitioners.
2. Educate researchers, who are interested or currently involved in medical imaging, in basic and advanced statistical techniques and make these techniques available through open-source software packages in *R*.

Potential Attendees

R users (clinicians, statisticians, medical physicists or researchers) with an interest in the quantitative analysis of neuroscience and/or oncology MRI data.

Background Knowledge

Attendees will require a basic understanding of an interpreted programming language; such as R (preferred) or Matlab. Attendees will also require basic understanding of statistical methodology; such as summary statistics, hypothesis tests, linear regression, non-linear regression, etc. Basic knowledge of medical imaging (specifically MRI) is an advantage but not necessary.

References

- [1] Polzehl, J. and K. Tabelow (2011). Beyond the Gaussian model in diffusion-weighted imaging: The package **dti**. *Journal of Statistical Software* 44(12), 1–26.
- [2] Tabelow, K., J. D. Clayden, P. Lafaye de Micheaux, J. Polzehl, V. J. Schmid, and B. Whitcher (2011). Image analysis and statistical inference in neuroimaging with R. *NeuroImage* 55(4), 1686–1693.
- [3] Tabelow, K. and J. Polzehl (2011). Statistical parametric maps for functional MRI experiments in R: The package **fmri**. *Journal of Statistical Software* 44(11), 1–21.
- [4] Tabelow, K. and B. Whitcher (2011). Special volume on magnetic resonance imaging in R. *Journal of Statistical Software* 44(1), 1–6.
- [5] Whitcher, B. (2011). Medical imaging task view. <http://cran.r-project.org/web/views/MedicalImaging.html>.
- [6] Whitcher, B. and V. J. Schmid (2011). Quantitative analysis of dynamic contrast-enhanced and diffusion-weighted magnetic resonance imaging for oncology in R. *Journal of Statistical Software* 44(5), 1–29.
- [7] Whitcher, B., V. J. Schmid, and A. Thorton (2011). Working with the DICOM and NIFTI data standards in R. *Journal of Statistical Software* 44(6), 1–29.

Tutorial: Emacs Speaks Statistics

Richard M. Heiberger^{1,*}, Martin Maechler²

1. Temple University, Philadelphia, PA

2. ETH Zurich

*Contact author: rmh@temple.edu

Keywords: ESS, R, development environment

Abstract

This tutorial will introduce the *emacs* environment, *ESS* (Emacs Speaks Statistics), for editing *R* code and interacting with *R*.

Emacs Speaks Statistics (ESS) is an add-on package for emacs text editors. It is designed to support editing of scripts and interaction with various statistical analysis programs such as R, S-Plus, SAS, Stata and JAGS. Although all users of these statistical analysis programs are welcome to apply ESS, advanced users or professionals who regularly work with text-based statistical analysis scripts, with various statistical languages/programs, or with different operating systems might benefit from it the most.

Goals

Show new users the advantages of using *ESS* and *emacs*. Show existing users newly-developed tools, such as those for literate programming (both **Sweave** and **Org-babel**).

Outline

1. Introduction (5 minutes)
2. Using Emacs (45 minutes)
3. Using ESS (60 minutes)
Exercise 1: ESS
4. Emacs extensions, including Sweave and Org-babel (40 minutes)
Exercise 2: Sweave
5. Emacs Lisp and customization (20 minutes)
6. Future directions for ESS, wrap-up (10 minutes)

Prerequisites

Attendees should have a working knowledge of *R*. Some previous experience with *Emacs* is helpful, but not essential. Bring your laptop with you to complete the exercises. Please install *R* on it, and if possible, *Emacs* and *ESS*. Precompiled *Emacs+ ESS* bundles for Windows and Mac are available on the *ESS* home page <http://ess.r-project.org/>.

Intended Audience

We hope to attract both new users and provide a base for existing users to swap ideas and discuss future plans for ESS.

Material for Participants

Some material from the User! 2011 ESS Tutorial is available <http://www.damtp.cam.ac.uk/user/sje30/ess11/>. This year's material will be available in May 2012.

Tutorial: A Crash Course in R Programming

Olivia Lau

Google, Inc.
Olivia.Lau@Post.Harvard.Edu

Keywords: statistical computing, lexical scoping, programming, profiling

Abstract

This course will cover common problems in R code, and offer techniques to make code more efficient and less prone to user error. By the end of the course, participants should be able to write efficient statistical software and validate existing code.

Outline

Topics will include:

- 1) Overview of R (20 minutes)
 - a) Units: Numeric, integer, character, and factor
 - b) Data structures: Vector, matrix, array, list, and data frames
 - c) Manipulating data structures
- 2) Writing efficient R code (40 minutes)
 - a) Writing (and avoiding) loops
 - b) Vectorizing over matrices, arrays, and lists
- 3) Writing functions (20 minutes)
 - a) Environments
 - b) Passing arguments and proper lexical scoping
- 4) Making R code easy-to-use (40 minutes)
 - a) Implementing model functions in your code
 - b) Generic functions
- 5) Validating statistical models (30 minutes)
 - a) Monte Carlo simulation
 - b) Unit tests
- 6) Making R run fast (30 minutes)
 - a) Profiling R code
 - b) Calling compiled languages (C, C++, Fortran) from R

Intended Audience

UseRs (and programmers from other statistical languages) who are interested in learning how to write fast, easy-to-use R code.

Required Knowledge

Familiarity with a command line-based statistical package (e.g., R, SAS, etc) is required. Instructions on installing R and an R editor will be distributed via email prior to the course.

Disclaimer: The views expressed in this presentation are those of the presenter and must not be taken to represent policy, guidance, or an endorsement of any software on behalf of Google, Inc.

Creating effective visualisations

Hadley Wickham^{1*}

1. Department of Statistics, Rice University

*Contact author: hadley@rice.edu

Keywords: visualisation, visualization, statistical graphics.

This half-day course will help you create better visualisations by teaching you about the findings from cognitive psychology that help us understand how the brain processes visual information. You'll learn important principles that underlying all visual displays and find out about some common mistakes that lead to confusion and inaccurate perception. We'll focus on four important principles:

- Match perceptual and data topology
- Make important comparisons easy
- Visual connections should reflect real connections
- Beware of animation!

The class will be a mixture of lecture and small-group activities, where you'll apply your new skills to critique existing visualisations and suggest improvements. Bring along one or two visualisations that you've been struggling with.

While some of the examples will use the ggplot2 R package, this course is graphics package agnostic. You'll be able to apply the skills you learn to any visualisation task, whether its in R or in another environment.

Getting the Most Out of RStudio™

Josh Paulson^{1,*}, JJ Allaire¹, Joe Cheng¹

1. RStudio

*Contact author: josh@rstudio.org

Keywords: RStudio, IDE, Web, Tools

Abstract:

RStudio is a free and open source integrated development environment (IDE) for the *R* programming language. RStudio aims to combine the various components of *R* (console, source, editing, graphics, history, help, etc.) into one seamless and productive workbench. RStudio also makes it easier to separate your workflow into projects and utilize other powerful tools such as version control. It is designed to both ease the learning curve for new *R* users as well as provide high productivity tools for more advanced users. Additionally, RStudio can be deployed as a server to enable web access to *R* sessions running on remote systems. With all of these powerful tools, it is important to know how to use them efficiently to maximize productivity.

For useR! 2012, we would like to give a full tutorial on getting the most out of RStudio. We will provide a full overview of the IDE, as well as instruction on setting up and using more advanced features such as the **manipulate** package, version control, and RStudio Server. Below is an outline of the proposed session:

Outline:

- Introduction
- RStudio Basics
- Full Feature Overview
- Productivity Techniques
- **manipulate** Package
- Integrated Version Control with Git and Subversion
- Setting up RStudio Server

Prerequisites: Attendees should have a basic working knowledge of *R*. Bring a laptop, preferably with *R* and the latest version of RStudio installed. RStudio binaries are available for Windows, Mac, and Linux on the RStudio website.

Intended Audience: New and advanced *R* users. The focus of this tutorial will be how RStudio can enhance your workflow and productivity while using *R*.

References:

[1] RStudio, Inc. (2012). RStudio home page, <http://www.rstudio.org/>.

Tutorial: Advanced Rcpp Usage

Dirk Eddelbuettel^{1,*}, Romain François^{2,*}

1. Debian Project
2. R Enthusiasts

*Contact both authors: RomainAndDirk@r-enthusiasts.com

Keywords: R, C++, Programming, Interfaces

Topic: This tutorial will provide a hands-on introduction to more advanced **Rcpp** features.

We intend to cover topics such as

- writing packages that use **Rcpp**,
- how to use *Rcpp modules*, and how the *R, ReferenceClasses* interact with *Rcpp modules*
- how *Rcpp sugar* lets us write C++ code that is close to R code in its expressiveness and use of implicit vectorisation, yet runs at the speed of compiled code,
- using the **RInside** package to embed R code in C++ applications.

Prerequisites: Knowledge of R as well as general programming knowledge; prior C++ knowledge may be helpful as well.

Equipment: Users should bring a laptop set up so that R packages can be built. That means on Windows, the **Rtools** bundle needs to be present and working, on OS X the **Xcode** package should be installed, and on Linux things generally just work. We can provide limited assistance in getting the required tools installed but the focus of the tutorial on how to use them.

References

- [1] Eddelbuettel, D. and R. François (2010). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40(8), 1–18.
- [2] Eddelbuettel, D. and R. François (2011). *Rcpp: Seamless R and C++ Integration*. R package version 0.9.9.
- [3] Eddelbuettel, D. and R. François (2012). *RInside: C++ classes to embed R in C++ applications*. R package version 0.2.6.

Design of the Survival Packages

Terry Therneau*

1. Mayo Clinic

*Contact author: therneau.terry@mayo.edu

Keywords: packages, survival

Half day tutorial proposal

In 1984 I created the first components of the survival package, which endures as one of the suggested components of R. In the interim 5 other packages have been created: date (deprecated), rpart, kinship, bdsmatrix, and coxme. What have I learned along the way? Much of this is captured in the layout of the final 2 packages, one of which uses S4 and one S3 type methods.

In course will present design/implementation issues that arise in the design of a package along with the choices that were made in the survival/coxme packages, why those were chosen, and the consequences in terms of what worked well or not so well. The primary document for the class will be a guide (currently in progress), which will allow us to focus on higher level concepts while leaving the details for home. Attendees should be able to use the course and the guide to write their own packages and not overlook important facets.

Some areas to be covered are

- Top level design: goals of the function, arguments, outputs, one package or multiple interlinked ones.
- Infrastructure: Rforge, CRAN, package skeletons, integrated documentation such as noweb (Sweave) or roxygen.
- Modeling functions: formula processing, data set handling, special variables and transformations, and print/summary/residual/predict methods
- Objects: Different kinds of objects including model fits, survival curves, pedigrees, and special matrices. What components do they have and why, and tradeoffs between rigidity and flexibility.
- Documentation and test suites

Bioconductor for High-Throughput Sequence Analysis

Martin Morgan^{1,*}

1. Fred Hutchinson Cancer Research Center, Seattle, WA

*Contact author: mtmorgan@fhcrc.org

Keywords: Bioinformatics, DNA Sequence, RNA-seq, ChIP-seq

DNA sequence analysis generates large volumes of data presenting challenging bioinformatic and statistical problems. This tutorial introduces *Bioconductor* packages and work flows for the analysis of sequence data. We learn about approaches for efficiently manipulating sequences and alignments, and introduce common work flows and the unique statistical challenges associated with ‘RNA-seq’, ‘ChIP-seq’ and variant annotation experiments. The emphasis is on exploratory analysis, and the analysis of designed experiments. The workshop assumes an intermediate level of familiarity with R, and basic understanding of biological and technological aspects of high-throughput sequence analysis. The workshop emphasizes orientation within the *Bioconductor* milieu; we will touch on the **Biostrings**, **ShortRead**, **GenomicRanges**, **edgeR**, and **DiffBind** packages, with short exercises to illustrate the functionality of each package. Participants should come prepared with a modern laptop with current *R* installed.

Predictive Modeling with *R* and the caret Package

Max Kuhn¹

1. Nonclinical Statistics, Pfizer Global R&D

*Contact author: max.kuhn@Pfizer.com

Keywords: Machine Learning, Pattern Recognition, Classification, Regression

This course will provide an overview of using *R* for supervised learning. The session will step through the process of building, visualizing, testing and comparing models that are focused on prediction. The goal of the course is to provide a thorough workflow in *R* that can be used with many different modeling techniques. A case study is used to illustrate functionality.

Outline:

- Introduction (philosophy, case study)
- General Strategies (data splitting, resampling, model tuning)
- Data PreProcessing (transformations, variable filtering)
- Conventions in *R* (OOP, function interfaces, consistency)
- Building and Tuning Models (performance metrics, trees, kernel methods, comparing models)
- Other Topics (as time allows) (parallel processing, variable importance)

Required Background Knowledge

Basic understanding of *R* (matrices, data frames, functions, etc) is needed. Some basic understanding of regression techniques is helpful.

Managing Data with *R*

Robert A. Muenchen*

The University of Tennessee

*Contact author: muenchen@utk.edu

Keywords: data manage transform reshape join

When analyzing a typical data set, we often spend the most of our time preparing the data. We join files adding variables or observations, transform the data using formulas, recode values, manage missing values, reshape the data from “wide” to “long” and create aggregate data sets by group(s) to use directly or to merge back to the original one. This half-day workshop will show you how to do these and other standard steps in R, pointing out common problems and how to solve them. Participants are encouraged to bring a laptop computer. The practice data and program will be online the week before the workshop at <http://r4stats.com>.

Tutorial: Geospatial Data in R and Beyond

Barry Rowlingson^{1,*}

1. Faculty of Health and Medicine, Lancaster University

*Contact author: b.rowlingson@lancaster.ac.uk

Keywords: geospatial, gis, maps

Overview

Spatial data is, quite literally, everywhere. In the past it was the private property of the GIS lab, but now everyone seems to be making tracks with the GPS device in their pocket. Map data now extends from the personal to the political, as agencies and governments make more global information available to the public. In recent years *R* has become well equipped to deal with this spatial data deluge, with a number of packages dedicated to spatial data and spatial analyses.

Previous tutorials at UseR! meetings have concentrated on statistical analysis of spatial and spatial-temporal data. This tutorial will get back to basics in a way, and examine the issues involved in dealing with data on the map.

Goals

By the end of the session the participants will know about: the different types of spatial data; reading and writing data in various formats; manipulating and transforming spatial data; making maps with *R* **base** graphics functions; exporting to standard data formats; working with OGC standards; point, line, and polygon data frames with **sp**; raster data with **raster**; advanced geometric operations with **rgeos**. There will also be introductions to other GIS software and how they can work with *R* in synergy.

With its data-centric focus, this tutorial cuts across disciplines to be useful to anyone working with statistics in the real world.

Cloud Computing for the R environment

Karim Chine^{1,*}

1. Cloud Era Ltd

*Contact author: karim.chine@gmail.com

Keywords: Cloud computing, HPC, GUI, Web, Collaboration

The tutorial will introduce a new *R* package and a new visual GUI designer created to help *R* users in taking advantage of cloud computing and in leveraging the web to create and publish easily interactive applications and reports based on *R*'s capabilities.

The **elasticR** package makes it possible to use the Amazon cloud programmatically from a regular *R* session. *R* servers with a rich and stateful interface can be created on *EC2* with simple *R* functions and used to offload time-consuming computations to machines of large capacities, to apply *R* functions to large data sets in parallel and to collaborate in real-time.

RBoard is a virtual collaborative environment for creating in the cloud, visually or programmatically, User Interfaces and dashboards based on *R* functions and data. Widgets of various complexities (spreadsheet elements, sliders, *R* graphics viewers, regular, interactive and motions charts, *R* macros and data links, *Html 5* and *Java* Plugins, etc.) can be composed into virtual panels and published to the web like a *Google Document*.

Topics of this tutorial will include:

- An overview of cloud computing technologies and of the **Elastic-R** platform.
- Exercises to familiarize users with the most important functionalities of the **elasticR** package.
- Collaborative exercises to create and publish advanced *R*-based dashboards using **RBoard**.

Users are expected to have working familiarity with *R* and to have an up-to-date installation of *R*, *Java* and *Flash*.

Resources about the tutorial are available at the following address : <http://www.elastic-r.net/user2012>.

References

- [1] Karim Chine (2010). Elastic-R Platform, <http://www.elastic-r.net>.
- [2] Karim Chine (2010). Open science in the cloud: towards a universal platform for scientific and statistical computing. In: Furht B, Escalante A (eds) Handbook of cloud computing, Springer, USA, pp 453–474. ISBN 978-1-4419-6524-0.
- [3] Karim Chine (2010). Learning math and statistics on the cloud, towards an EC2-based Google Docs-like portal for teaching / learning collaboratively with *R* and Scilab, icalt, pp.752-753, 2010 10th IEEE International Conference on Advanced Learning Technologies.

Every Plot Must Tell a Story - even in *R*

Di Cook

Department of Statistics, Iowa State University
dicook@iastate.edu

A long, long time ago, in the early 1990's, when I first heard John, and Allan, and Rick talking about *S*, they were taking great care in the way they described the software, as

“the language of data analysis”,

ferociously at times, asserting that it should *not* be described as “statistical software”. This carefully guarded perception was successfully sustained when the the ACM awards committee acknowledged John Chambers contributions to computing as

“the *S* system, which has forever altered the way people analyze, visualize, and manipulate data ...”

Today *R* makes the *S* language readily accessible, and helps to disseminate data, albeit, often rudimentary data, that are used as examples of methods available in base *R* and contributed packages. In this talk we take a critical look at the way data are introduced, especially, graphically, in *R*. Graphics are good for showing the information in datasets and for complementing modeling. Sometimes graphics show information models miss, sometimes graphics help to make model results more understandable, sometimes models show whether information from graphics has statistical support or not. Above all, a good selection of graphics tells a story of the data analysis, and this could be better harnessed in *R*.

Joint work with Antony Unwin and Heike Hofmann

Simplified Reproducible Research using lazyWeave

Nutter Benjamin¹

1. Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio

Keywords: Reproducibility, \LaTeX , Sweave, lazyWeave

Sweave has been a major contributor to reproducible research in the R/S communities. However, it was built on the assumption that “Many S users are also \LaTeX users, hence no new software has to be learned [besides R/S][1].” Users new to R and/or \LaTeX , however, have steep learning curves to climb before being able to develop reproducible reports of their research and analyses.

lazyWeave provides an option to generate reproducible reports entirely within the R environment and with very little knowledge of \LaTeX . **lazyWeave** also provides a flexible framework in which to build complex and customized tables. Tables may be built using one or more matrix[-like] objects and may include horizontal lines, colored rows, and user-defined column widths.

Other features made available through **lazyWeave** are tables of contents, lists, counters, captions, labels, footnotes, page numbering, and page breaks. More advanced users may use **lazyWeave** to write functions that generate complete, production ready tables, figures, or entire reports with the use of a simple command. While not suitable as a replacement of **Sweave**, it does provide an intermediate step toward reproducible research for R users who aren't yet ready to learn \LaTeX .

References

- [1] Leisch, F. (2002). “Sweave: Dynamic generation of statistical reports using literate data analysis.”. In W. Hardle and B. Ronz (Eds.), *Compstat 2002 – Proceedings in Computational Statistics*, pp. 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9, URL <http://www.stat.uni-meunchen.de/~leisch/Sweave>.

Predicting Dangerous *E. coli* Levels at Erie, Pennsylvania Beaches Using *R*

Michael A. Rutter^{1,*}

1. Penn State Erie, The Behrend College

*Contact author: marutter@gmail.com

Keywords: Data mining, Water Quality, Data Collection, Web Scraping

Presque Isle State Park in Erie, Pennsylvania is home to 11 public beaches on Lake Erie that attract more than 4 million visits annually. During the summer swim season, water quality is always a public health concern, specifically the presence of *Escherichia coli* (*E. coli* for short). While *E. coli* itself can be dangerous to humans, it is also an indicator of other, harmful bacteria that are much harder to detect. In order to monitor the beaches, water samples are taken in the morning to measure *E. coli* levels and issue warnings or beach closings if necessary. While the protocol is well established, a major downside is that it takes 24 hours for the water samples to be tested, therefore *E. coli* levels are known only after swimming for the day has been completed.

Given the time limitations of the testing protocol, there is a need for a statistical model that attempts to predict *E. coli* levels before swimmers arrive at the park based on weather conditions and other data sources available on the internet. In this paper I will present some of the details of the development of a model to predict *E. coli* levels at Presque Isle beaches using a random forest decision tree approach. In addition, I will show how the entire process of collecting data from the internet, generating predictions from the model, and presenting the results in a web page are done completely within *R*. Of special interest is the relative ease in which *R* can obtain (or scrape) data from public web sites so real time data can be used in model predictions.

Package development: the easy way.

Hadley Wickham^{1*}

1. Department of Statistics, Rice University

*Contact author: hadley@rice.edu

Keywords: package development, devtools

In this talk, I'll outline my philosophy of package development, and show how the **devtools** package makes package development easy (even on windows!) and protects you from many of the problems I've suffered when developing packages. I'll discuss the following important components of the development cycle:

- Iteratively modifying your code and checking if it works
- Documenting functions with a minimum of duplication (using **roxygen2**)
- Developing automated test cases to ensure that your code works as expected (using **testthat**)
- Passing R CMD check.

I'll also touch on some of the other useful tools that devtools provides for installing remote packages, running remote code and checking that your development environment is correctly set up.

Applications of R in Business

David Smith^{1,*}

1. Revolution Analytics

*Contact author: david@revolutionanalytics.com

Keywords: R, applications, sentiment analysis, forecasting, manufacturing, clinical trials, reporting, supply chain optimization

The “Applications of R in Business Contest” was created by Revolution Analytics to find the best examples of applying R to business problems, and to create an on-line resource of tools to help users apply R to real-world problems. With \$20,000 in cash prizes, the contest attracted many compelling examples of R usage. In this talk, I’ll review the prizewinning applications, including: forecasting the performance of a marketing campaign; sentiment analysis of airlines using Twitter; predicting duration of clinical trials; monitoring temperature in steel production; estimating duration of IT projects; forecasting orders for supply chain optimization; and reporting using the **knitr** package. I’ll also give an update on how these applications are impacting businesses today, and give a preview of the next iteration of the contest.

References

- [1] *Revolution Analytics Announces "Applications of R in Business" Contest Winners* (press release, January 25, 2011), <http://www.revolutionanalytics.com/news-events/news-room.php>

Give Your Data a Listen

Eric Stone^{1,2,*}, Jesse Garisson^{3,#}

1. PhD Student, Temple University Department of Statistics
2. US Department of Agriculture, National Agricultural Statistics Service
3. Independent Multimedia Artist, Brooklyn, NY: www.takethefort.com

*Contact author: ericstone@temple.edu

#Contact author: jesse@takethefort.com

Keywords: Data Visualization, Sound, Audiolize, Accessibility

Many methods exist for visualizing data in *R*. This paper seeks to devise a method for going beyond data visualization in *R* by translating existing methods into sounds, scales, and chords that convey important characteristics of data and datasets aurally. We do this by utilizing the existing arsenal of data visualization tools in *R*, converting them into a semi-standardized form, and then compiling and playing them using the popular audio programming toolkit *Max/MSP*. We provide what we refer to as “audiolized” versions of visualization and analysis methods including but not necessarily limited to hex binning (**hexbin**), principal component analysis (**prcomp**), clustering (**biclust**), and factorization (**Hmisc**). There has been some work on this topic, most notably on “sonification” (Borasky, Edward [1]), but we are not aware of any such system for *R*. While the present project is targeted at users who are familiar with *R*, by using *RExcel*, our data audiolization can be made accessible to the large base of *Excel* users. Our current goals are (1) to provide an additional, secondary, method for observing data prior to further analysis and model-building; (2) to display results of data analysis aurally; (3) to engage visually impaired users with data and analyses in an innovative manner; and (4) make compelling, data-driven musical compositions.

References

- [1] Borasky, Edward (2012). Listening to Data: Sonification Using Open Source Tools, <http://opensourcebridge.org/sessions/376>.

Modelling the Effect of Social Mobility on Limiting Long-term Illness

Heather Turner^{1,2,*}, David Firth¹

1. University of Warwick, UK
2. Independent statistical/R consultant
*Contact author: ht@heatherturner.net

Keywords: Generalized nonlinear models, binary data, public health, census data

A change in a person's social class may have a number of effects on the person themselves or the way they live their life. Hence researchers have considered the effect of social mobility in relation to varied outcomes such as the risk of cancer [3], consumption of alcohol [1] and attitudes towards ethnic minorities [5].

An intuitive model for the effect of social mobility is the diagonal reference model proposed by Sobel [4]. In this model, the effect for individuals moving from origin class i to destination class j is defined as a weighted sum of the i 'th and j 'th diagonal effects:

$$w_1\gamma_i + (1 - w_1)\gamma_j,$$

where a diagonal effect is the effect for stable individuals in a given class. This model has predominantly been applied to the analysis of political and social attitudes, measured on a scale that can be assumed to follow a normal distribution [e.g. 5, 2]. In this case the diagonal reference model can be estimated using nonlinear least squares, for example with `nls`.

In this talk we use the diagonal reference model to investigate the effect of social mobility on limiting long-term illness, using data from the UK's Office of National Statistics Longitudinal Study. In this case the outcome is binary, hence the model is an example of a generalized nonlinear model, which can be fitted in R using the `gnm` package [6].

References

- [1] Hart, C. L., G. Davey Smith, M. N. Upton, and G. C. M. Watt (2009). Alcohol consumption behaviours and social mobility in men and women of the Midspan Family study. *Alcohol and alcoholism* 44(3), 332–6.
- [2] Paterson, L. (2008). Political attitudes, social participation and social mobility: a longitudinal analysis. *British Journal of Sociology* 59(3), 413–434.
- [3] Schmeisser, N., D. I. Conway, P. A. McKinney, A. D. McMahon, H. Pohlabein, M. Marron, S. Benhamou, C. Bouchardy, G. J. Macfarlane, T. V. Macfarlane, P. Lagiou, A. Lagiou, V. Bencko, I. Holcátová, F. Merletti, L. Richiardi, K. Kjaerheim, A. Agudo, R. Talamini, J. Polesel, C. Canova, L. Simonato, R. Lowry, A. Znaor, C. Healy, B. E. McCarten, M. Hashibe, P. Brennan, and W. Ahrens (2010). Life course social mobility and risk of upper aerodigestive tract cancer in men. *European journal of epidemiology* 25(3), 173–82.
- [4] Sobel, M. E. (1981). Diagonal mobility models: A substantively motivated class of designs for the analysis of mobility effects. *Amer. Soc. Rev.* 46, 893–906.
- [5] Tolsma, J., N. D. de Graaf, and L. Quillian (2009). Does intergenerational social mobility affect antagonistic attitudes towards ethnic minorities? *The British journal of sociology* 60(2), 257–77.
- [6] Turner, H. and D. Firth (2011). *Generalized nonlinear models in R: An overview of the gnm package*. R package version 1.0-1.

Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse

Daniel W. Connolly^{1*}, Bhargav Adagarla¹, John Keighley¹, Lemuel R. Waitman¹

1. Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas

*Contact author: dconnolly@kumc.edu

Keywords: Data Warehouse, Security, Biomedical Informatics

HERON[1] is a clinical data repository with about a half a billion facts, linking hospital medical records and clinic billing systems with our tumor registry and biospecimen repository and national databases such as the Social Security death index. It is built on i2b2, an NIH-funded scalable informatics framework[2]. Users can interactively answer questions such as “how many patients meet my study criteria?” and use analysis plug-ins to visualize the potential study cohort’s demographic characteristics. Bulk export of data for off-line analysis is allowed but only after explicit approval by a governance structure that safeguards patient privacy. We aim to package a range of statistical analysis methods for secure, on-line analysis within i2b2.

On top of a star-schema database and a middle tier web services *hive of cells*, i2b2 provides a web based user interface. Previous work developed *RECell* [3] to integrate R into the i2b2 architecture and a Kaplan Meier analysis plug-in that relays patient data as XML to the *RECell*. The *RECell* then transforms data as required by the R **survival** package, and invokes R to produce a plot which is fetched and displayed by the plug-in. Rather than sending all the data via the web client and serializing/parsing it several times, our approach, *rgate*, connects R directly to the database, using the **DBI** and **ROracle** packages, performing the data transformation in R, and proceeds with the survival package as above.

Authority flows from an oversight committee (hospital, clinics, and university), which grants users authority to view data and grant operators authority to run the service; the operator grants *rgate* authority to query the database via a configuration file; users present login credentials as proof of their authority to view data; the plug-in relays this authority to *rgate* in the form of a session identifier.

The R code invoked by *rgate* is split between a trusted deid.R broker and an untrusted analysis.R module. The trusted broker module has a function that takes (1) database login credentials and (2) a patient set identifier and returns an facet object[4] that will only query attributes of those patients. Only this facet object is given to the untrusted analysis module. This isolates the statistician from security and governance issues to ensure their statistical analysis modules do not cause users to exceed the authority given to them by the oversight committee.

References

[1] Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing Observations from Electronic Medical Record Flowsheets in an i2b2 based Clinical Data Repository to Support Research and Quality Improvement. AMIA Annu Symp Proc. 2011;2011:1454-63. Epub 2011 Oct 22.

[2] Murphy SN et al. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc.* 17, 124–130.

[3] Segagni D, e. a. (2011, May). R engine cell: integrating R into the i2b2 software infrastructure. *J Am Med Inform Assoc.* 18(3), 314–7.

[4] Miller, M. S. (2006). Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. <http://www.erights.org/talks/thesis/>.

Gene Expression Analysis In Red-R

Kyle R. Covington^{1,2,*}, Suzanne A. W. Fuqua¹

1. Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030

2. Red-R.org

*Contact author: kyle@red-r.org

Keywords: Gene Expression, Bioinformatics, Visual Programming

Hundreds of published gene expression data-sets are publically available. These data-sets continue to be mined for informative relationships between gene expression and biological or clinical markers. Several tools have been developed to facilitate public access to these data-sets, such as Oncomine [2]. Unfortunately, these tools often limit users in the types of comparisons that they can perform and do not expose the underlying analysis engine to advanced users. More powerful analysis tools, such as *R*, are often difficult for novice users to use.

To address these problems, we have developed Red-R [1], a graphical user interface for *R*. With this tool, we are able to leverage the statistical powers of *R* in an environment in which analysis pipelines can be rapidly generated. Pipelines refresh data across the program so that data remains accurate as the user interacts with the data. Pipelines can be extended at any time by adding additional analysis nodes (widgets). Pipelines can be saved and reloaded so that all data and analysis methods are linked together so that analyses are highly reproducible. In this way, users can quickly generate custom programs that perform specific tasks customized to the users specific needs. As a demonstration of this, we show that Red-R can be used to generate pipelines to study the association of biological and clinical data across several published data-sets using tools for Affymetrix microarray, survival, and differential gene expression analysis. More information on this project is available at <http://www.red-r.org>.

References

- [1] Covington, K. R. and A. Parikh (2011, August). The red-r framework for integrated discovery. *The Red-R Journal* 1-08/08/2011.
- [2] Rhodes, D. R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan (2004). Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6(1), 1–6.

R package *ROCS* for Inferring Specific Protein-Protein Interactions in AP-MS Data

Jean-Eudes Dazard^{1,*}, Sudipto Saha¹, Robert Ewing¹

1. Division of Bioinformatics, Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, Ohio

*Contact author: jxd101@case.edu

Keywords: R package, Experimental Reproducibility, Protein-Protein Interaction, Affinity-Purification Mass-Spectrometry, Parallel Programming.

Affinity-Purification Mass-Spectrometry (AP-MS) provides a powerful means of identifying protein complexes and interactions. Several important challenges exist in interpreting the results of AP-MS experiments, including the reproducibility between AP-MS experimental replicates due both to technical variability and the dynamic nature of protein interactions in the cell; and the accuracy in the identification of true protein-protein interactions due to high false negative and false positive rates. To address these issues, we recently introduced a two-step method, called *ROCS*, which makes use of *Indicator Proteins* to select reproducible AP-MS experiments, and of *Confidence Scores* to select specific protein-protein interactions [1]. Our *Indicator Proteins* account for measures of protein identifiability as well as protein reproducibility, effectively allowing removal of outlier experiments that bring noise and affect further inferences. The curated set of experiments is then used in the Protein-Protein Interaction (PPI) scoring step. Actual prey protein scoring is done by computing our *Confidence Score*, which accounts for the probability of occurrence of prey proteins in the bait experiment relative to the control experiment, where the significance cutoff parameter is estimated by simultaneously controlling false positives and false negatives against metrics of false discovery rate as well as biological coherence respectively. Altogether, the *ROCS* two-step approach relies on automatic objective criteria for parameter estimation and error-controlled procedures. We show that our method may be used on its own to make accurate identification of specific, biologically relevant protein-protein interactions or in combination with other AP-MS scoring methods to significantly improve inferences. We present an implementation of this recent method in the *R* language for statistical computing. The *R* package offers a complete implementation including: (i) computation of *Indicator Proteins*, and Protein-Protein Interaction (PPI) *Confidence Scores* (ii) interactive estimation of parameters by FDR-controlled and Gene Ontology semantic similarity procedures, (iii) generation of diverse diagnostic and error plots, (iv) option of efficient parallel computation by means of the *R* package **snow**, (v) examples by real AP-MS experiment datasets, each containing well characterized interactions, allowing for systematic benchmarking of *ROCS*, (vi) a manual and documentation. To make each feature as user-friendly as possible, only one wrapper subroutine per functionality is to be handled by the end-user. It is available as an *R* package, called 'Reproducibility Index and Confidence Score' ('**ROCS**'), downloadable from the CRAN website.

References

- [1] Jean-Eudes Dazard, Sudipto Saha, Robert Ewing (2012). *ROCS: A Reproducibility Index and Confidence Score for Improved Identification of Specific Protein-Protein Interactions in Affinity-Purification Mass-Spectrometry Data* *BMC Bioinformatics* (in revision).

Knowledge Discovery in Neglected Disease Databases: Using *R* for Data Visualization, Analysis and Mining

Paul J Kowalczyk, PhD

SCYNEXIS, P O Box 12878, Research Triangle Park, NC 27709-2878

paul.kowalczyk@scynexis.com

Keywords: distributed models, machine learning, neglected diseases

Active collaborations between SCYNEXIS and DNDi[1], TDR[2], and MMV[3] have identified both opportunities and challenges in building and deploying a cheminformatics toolkit for knowledge discovery in neglected diseases databases. We describe our use of *R* in the development of data visualization, data analysis and data mining work flows. Applications range from straightforward data summaries (*e.g.*, molecular weight distributions, number of H-bond donors, log P ranges, etc.) to pairwise comparisons of multiple databases to the construction of predictive machine learning models. In particular, we describe an approach to continuous virtual screening. This workflow makes use of new data, in real time, to construct, validate and deploy models automatically. The toolkit has the ability to build both categorization and regression models, as the data dictate. We show use cases for these workflows with open-source data for anti-malarial screening collections.

References

[1] DNDi: Drugs for Neglected Diseases *initiative*. <http://www.dndi.org/>

[2] TDR: the Special Programme for Research and Training in Tropical Diseases.
<http://www.who.int/tdr/>

[3] MMV: Medicines for Malaria Venture. <http://www.mmv.org/>

SVGMapping: an R package to map *omic* data sets onto pathways templates

Raphaël Champeimont^{1,2}, Christophe Leplat¹, Franck Chauvat¹, Jean-Christophe Aude^{1,*}

1. CEA, iBiTecS, Integrative Biology Laboratory, F-91191 Gif-sur-Yvette, France

2. Génomique des Microorganismes, UMR 7238 CNRS-UPMC, 15 Rue de l'École de Médecine, 75006 Paris, France

*Contact author: jean-christophe.aude@cea.fr

Keywords: Vizualisation, Pathways, Microarrays, High throughput assays

High-throughput *omic* technologies are now commonly used in large-scale experimental biology. The main characteristic of these *omic* approaches is that they usually produce large amounts of data. Results obtained through these analyses are mostly interpreted or assessed in terms of given hypotheses. In most cases, huge amount of results need to be transformed (*eg* using classification methods), integrated with other biological knowledge (*eg* pathways), and explored using mainstream or dedicated visualisation tools. Then, they can be meaningfully interpreted by biologists. Visualisation is crucial for an optimal understanding of the results emerging from the concerted analysis of shared material between experimental and computational researchers.

Directed visualisation methods [2] use *prior* knowledge in their process. In biology this knowledge is often depicted by networks. For example, Momin et al. [3] designed a method that combines a visualisation method and a prediction process to map transcriptomic data with predicted metabolite pools into pathways. Here we report **SVGMapping**, an R package to map *omic* experimental data onto custom-made templates which can be used to depict metabolic pathways, cellular structures or biological processes. **SVGMapping** [1] allows the modification of color, opacity or shape of given graphical elements. It can be applied several times on the same template to combine various *omic* data types (*eg* protein and metabolite concentrations). This package has been designed to integrate the wealth of data generated by various strains (*eg* mutants *vs* wild-type), growth conditions (*eg* before *vs* after stress) or kinetic experiments. Both templates and output graphics comply with the Scalable Vector Graphics (SVG) format. This format can be converted in popular graphic formats (*eg* PNG or PDF) or displayed within a modern browser. Using the latter, **SVGMapping** can be setup to encapsulate annotations as *tooltips* and *hyperlinks* to provide an interactive user-friendly experience.

References

- [1] Aude, J.-C. and R. Champeimont (2007). Svgmapping homepage. <http://svgmapping.r-forge.r-project.org/>.
- [2] Keim, D. A., F. Mansmann, J. Schneidewind, and H. Ziegler (2006, jul). Challenges in Visual Data Analysis. In *Tenth International Conference on Information Visualisation (IV'06)*, University of Konstanz, Germany, pp. 9–16. University of Konstanz, Germany: IEEE.
- [3] Momin, A. A., H. Park, B. J. Portz, C. A. Haynes, R. L. Shaner, S. L. Kelly, I. K. Jordan, and A. H. Merrill (2011, may). A method for visualization of "omic" datasets for sphingolipid metabolism to predict potentially interesting differences. *The Journal of Lipid Research* 52(6), 1073–1083.

ggmap : Interfacing ggplot2 and RgoogleMaps

David Kahle^{1,*}, Hadley Wickham²

1. Assistant Professor of Statistical Science, Baylor University

2. Assistant Professor of Statistics, Rice University

*Contact author: david.kahle@baylor.edu

Keywords: ggplot2, Google Maps, OpenStreetMap, layered grammar of graphics, spatial statistics

In spatial statistics the ability to visualize data and models superimposed with their basic social landmarks and geographic context is invaluable. **ggmap** is a new tool which enables just that by combining the spatial information of Google Maps or OpenStreetMap from **RgoogleMaps** with the layered grammar of graphics implementation of **ggplot2**. In addition to the full range of **ggplot2** features such as bubble/contour plots and faceting – plotted on top of maps – **ggmap** provides the following features

1. geocoding and reverse geocoding using the Google Maps API with lazy specification (e.g. `geocode("baylor university")`),
2. distance and travel time lookup via the Google Maps Distance Matrix API (for driving, bicycling, or walking), and
3. a `locator` function for **ggplot2** graphics called `gglocator`, which is particularly useful for spatial graphics.

The result is a convenient, consistent and modular framework for spatial graphics.

References

- P. Murrell (2011). Raster Images in R Graphics. *The R Journal*, Vol. 3/1, 48–54.
- M. Loecher and Sense Networks. RgoogleMaps: Overlays on Google map tiles in R., R package version 1.1.9.3. <http://CRAN.R-project.org/package=RgoogleMaps>.
- H. Wickham (2009). *ggplot2: elegant graphics for data analysis*. Springer, New York.
- H. Wickham (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28.
- L. Wilkinson (2005). *The Grammar of Graphics*, 2nd ed., Springer, New York.

Interactive Hammock Plots for Visualizing Categorical Data

Marie Vendettuoli^{1,2,*}, Heike Hofmann^{1,2,3}

1. Bioinformatics and Computational Biology Program

2. Human Computer Interaction Program

3. Department of Statistics

Iowa State University, Ames IA 50010

*Contact author: mariev@iastate.edu

Keywords: interactive graphics, visualization, categorical data, exploratory data analysis

Hammock plots are designed for visualizing categorical data and have been described as “a generalization of parallel coordinate plots” that addresses issues of overplotting and missing data, both situations where traditional parallel coordinate plots face difficulty. We present `qhammock`, an interactive implementation that depends on *R* packages `qtbase` and `qtpaint`. This interactive plot supports both querying and brushing, reducing the cognitive load necessary during exploratory data analysis. We also present a new axis display that extends the hammock plot to include information regarding sample size and/or conditional probabilities. For complex datasets, `qhammock` also supports linking to other plots created using the developmental package `cranvas`. We compare the effectiveness of `qhammock` to a traditional circle graph via a user study.

References

- Schonlau M (2003) Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. In Proceedings of the Section on Statistical Graphics, American Statistical Association.
- Aumann Y, Feldman R, Ben Yehuda Y, Landau D, Liphstat O, Schler Y (1999) Circle Graphs: New Visualization Tools for Text-Mining. In Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science

cranvas: Interactive statistical graphics in R based on Qt

Yihui Xie^{1,*}, Xiaoyue Cheng¹, Di Cook¹, Heike Hofmann¹

1. Department of Statistics, Iowa State University

*Contact author: xie@yihui.name

Keywords: interactive graphics, R language, Qt, data pipeline

Interactive graphics progresses us beyond the limitations of static statistical displays, particularly for exploring data, and analyzing models. A long missing feature in *R* graphics systems (either base or grid graphics) is support for interactivity. A number of standalone systems exist, GGobi, MANET and Mondrian, which support interactive displays of multivariate data, but they lack the extensibility, and tight integration with modeling that *R* furnishes. There have been several attempts to provide interactive graphics in *R*, from `locate()` (ages ago!) and `getGraphicsEvent()` in base *R*, to packages **RGtk**, **RGtk2**, **rggobi**, **teltk** and **iplots**. Iplots, the most recent, provides extensive interactive graphics, using the Java backend. It is fast and includes most common types of plots; basic operations in interactive graphics such as selection and zooming are also supported. However, it lacks some features such as the tour or direct support for color palettes.

In this talk, I will introduce a new *R* package, **cranvas**, which is based on several other packages to make fast (Qt), flexible (*R*) and elegant interactive statistical graphics in *R*. The drawing is based on two packages **qtbase** and **qtpaint**, which provide API's to the Qt libraries in *R*. The data structure is based on **plumbr** and **objectSignals**, which bring forward a new data structure called “mutable data”. These packages set up signals and listeners on data (implemented purely with *R*), so that changes in data can trigger changes in plots. The mutable data is also the foundation of a “data pipeline” behind **cranvas**, where events of statistical analysis importance, like variable transformation or dimension reduction, can be propagated through to the displays. The **cranvas** package aims to borrow from the design of **ggplot2**, which is based on a grammar of graphics (Wilkinson, 2005). Currently this package includes common statistical graphics, histogram, scatterplot, bar plot, boxplot, parallel coordinate plot and map, and tours, and common interactions, brushing, identifying, deletion, zooming, panning and different types of linking. Color palettes are supported. This talk will show examples of the usage, how the ideas relate to other interactive graphics work, what is under the hood, and what we are planning in the future.

References

Xie et al. (2012). **cranvas**: Interactive statistical graphics based on Qt. URL: <https://github.com/ggobi/cranvas>.

bespoke: A package to custom-make online web applications of statistics homework and practice problems

G. Brooke Anderson^{1*} and Roger D. Peng¹

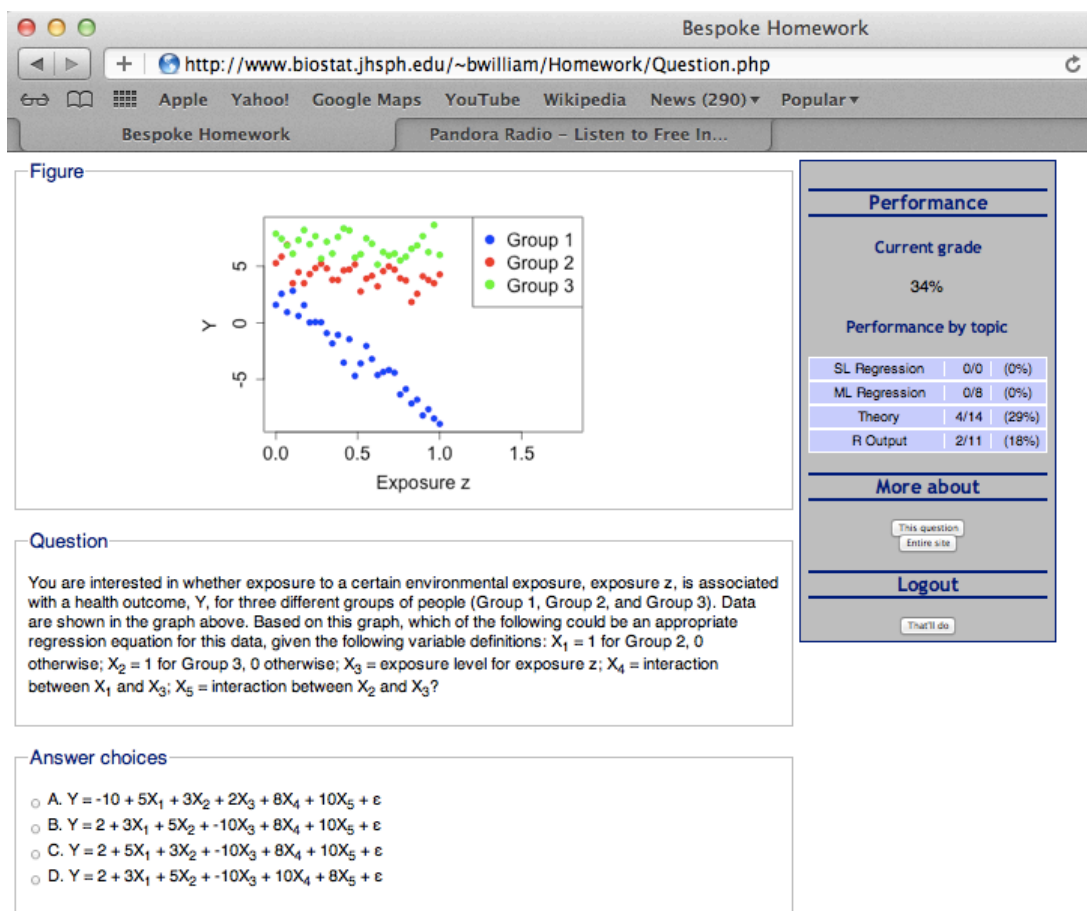
1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

*Contact author: geanders@jhsph.edu

Keywords: statistics education, automatic problem generation, multiple choice, package **bespoke**

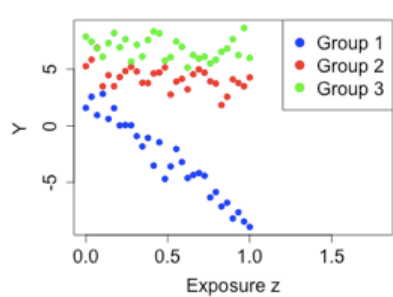
We are developing the **bespoke** package, which allows teachers to use R to custom-make web applications with course-specific multiple choice questions automatically generated from user-created question templates. One **bespoke** function takes user-created question templates, which include random numbers or randomly generated datasets as part of the input, and creates many versions of each question, with different numerical answers for each version. Another function inputs all automatically-generated versions of each question into a framework of HTML and PHP code, creating an online web application that can be loaded onto a server to provide students homework or extra practice (example screenshot shown below). Web applications created using the **bespoke** package include student-specific login and automatic grading of questions. We will describe use of this package and also share experiences from a pilot test of a **bespoke**-created homework web application in a Master's Level Biostatistics course at Johns Hopkins Bloomberg School of Public Health.

Figure: Screenshot of a homework web application created using the **bespoke** package.



The screenshot shows a web browser window with the URL <http://www.biostat.jhsph.edu/~bwilliam/Homework/Question.php>. The page is titled "Bespoke Homework" and contains a sidebar on the right with performance metrics and a main content area with a figure, question, and answer choices.

Figure



The scatter plot shows the relationship between Exposure z (x-axis, 0.0 to 1.5) and a health outcome Y (y-axis, -5 to 5). Group 1 (blue) shows a negative linear relationship. Group 2 (red) shows a positive linear relationship. Group 3 (green) shows a positive linear relationship with a steeper slope than Group 2.

Question

You are interested in whether exposure to a certain environmental exposure, exposure z , is associated with a health outcome, Y , for three different groups of people (Group 1, Group 2, and Group 3). Data are shown in the graph above. Based on this graph, which of the following could be an appropriate regression equation for this data, given the following variable definitions: $X_1 = 1$ for Group 2, 0 otherwise; $X_2 = 1$ for Group 3, 0 otherwise; $X_3 = \text{exposure level for exposure } z$; $X_4 = \text{interaction between } X_1 \text{ and } X_3$; $X_5 = \text{interaction between } X_2 \text{ and } X_3$?

Answer choices

- A. $Y = -10 + 5X_1 + 3X_2 + 2X_3 + 8X_4 + 10X_5 + \epsilon$
- B. $Y = 2 + 3X_1 + 5X_2 + -10X_3 + 8X_4 + 10X_5 + \epsilon$
- C. $Y = 2 + 5X_1 + 3X_2 + -10X_3 + 8X_4 + 10X_5 + \epsilon$
- D. $Y = 2 + 3X_1 + 5X_2 + -10X_3 + 10X_4 + 8X_5 + \epsilon$

Performance

Current grade: 34%

Performance by topic

SL Regression	0/0	(0%)
ML Regression	0/8	(0%)
Theory	4/14	(29%)
R Output	2/11	(18%)

More about

[This question](#)
[Entire site](#)

Logout

[That'll do](#)

Integrating R into introductory statistics

Mine Çetinkaya-Rundel^{1*}, Andrew Bray²

1. Duke University, Department of Statistical Science

2. UCLA, Department of Statistics

*Contact author: mine@stat.duke.edu

Keywords: introductory statistics, teaching, RStudio

In this talk we discuss approaches for effectively integrating *R* into an introductory statistics curriculum. *R* is attractive because, unlike software designed specifically for courses at this level, it is relevant beyond the introductory statistics classroom, and is more powerful and flexible. The main obstacle to the adoption and use of *R* in an introductory setting is the perceived challenge of teaching programming in addition to teaching statistical concepts. Furthermore, working at a command line tends to be more intimidating to students than more traditional GUI-based tools. Many of these challenges can be overcome with the right tools: a user-friendly IDE like RStudio, which is invaluable for resolving some of the initial hurdles novice students experience with the bare-bones *R* interface, and labs and activities that use the right balance of standard and custom functions. We will present examples from labs that have been developed with the goal of helping students synthesize concepts and apply them to real data. We will also share student experiences and discuss approaches for preparing teaching assistants to lead *R* based lab sessions.

Teaching R to large first year classes

Colin S. Gillespie

School of Mathematics & Statistics, Newcastle University, UK.
Contact author: colin.gillespie@ncl.ac.uk

Keywords: Large class teaching, applied statistics.

Around ten years ago, the School of Mathematics & Statistics at Newcastle University incorporated R into all of its undergraduate statistics courses. Prior to this, a variety of other statistical packages, such as Minitab and S-PLUS, were used. One of the goals of this switch was for students to develop a detailed appreciation of a single system, rather than a superficial understanding of multiple packages.

In stage one of their Mathematics degree, students take an introductory computational statistics course. This is a large class; there are typically around one hundred forty students. The goal of this module is not only to cover R programming, but also cover topics such as, random number generation and kernel density plots.

There are two main challenges with this module. First, the majority of class (initially) have a strong aversion to computing. Second, there is almost weekly coursework for this module. This coursework has to be marked and returned promptly.

This talk discusses strategies for overcoming these problems. To engage students, the module incorporates interesting Monte-Carlo simulation projects. For example, students simulate a basic Monopoly game, which would be mathematically intractable, but is amenable to a Monte-Carlo approach. To cope with the large marking burden, web-based assignments are used. Students upload their work, via web-forms, and personalised sweave reports are generated for each student.

Randomization methods in introductory statistics courses

Edith Seier

Department of Mathematics and Statistics

East Tennessee State University

In recent years, randomization methods such as permutation tests and the bootstrap are slowly but surely becoming part of introductory statistics courses. User-friendly applets are sometimes used to apply these methods. There are already packages in R that have functions to perform the randomization tests or bootstrapping. In both cases students can apply the procedures but might not be aware of what the computer is doing. We prefer to use a different approach: hands-on activities that facilitate the understanding of the process and programs in R specifically written to mimic those tactile experiences. This approach allows students to understand the procedure and 'see' that the software simply repeats the same procedure a large number of times. For example, after explaining the rationale of the randomization test to compare two populations, we do a hands-on experience with chips. We then ask the students to list the sequence of steps that were performed and then show them the program. This also gives the opportunity to mention that an algorithm is just a sequence of steps to perform a task and that writing a program consists in translating the algorithm into a language that the computer can understand. The programs are given to the students in a text file and they have simply to change the data and copy and paste the commands into R when solving exercises.

The poster presents three cases of tactile experiences coupled with our programs in R that we use in some introductory statistics courses. These are: the randomization test to compare two populations, the paired-data randomization test and percentile bootstrap confidence intervals.

Reference

Seier, E and Joplin, K. (2011) Introduction to Statistics in a Biological Context. CreateSpace.

msSurv, an *R* Package for Nonparametric Estimation of Multistate Models

A. Nicole Ferguson^{1,2,*}, Somnath Datta¹, Guy Brock¹

1. University of Louisville
2. Kennesaw State University

*Contact author: nicole.ferguson@kennesaw.edu

Keywords: survival, *R* package, nonparametric estimation, multistate models

We present an *R* package, **msSurv**, to calculate nonparametric estimates of the transition probability matrix, marginal state occupation probabilities, the normalized and non-normalized state entry and exit time distributions, and marginal integrated transition hazards for a general multistate system under left-truncation and right censoring. Excepting the transition probability matrix, the latter quantities have been shown to be valid even for non-Markov systems. State-dependent censoring is handled via modeling the censoring hazard conditional on the state currently occupied. Pointwise confidence intervals for the above mentioned quantities are obtained and returned using closed-form variance estimators for independent censoring and using the bootstrap for dependent censoring.

A Pipeline for Analysis of SNP Arrays for Longitudinal Studies (PASALS)

Laura Tipton^{1*}, David Altshuler^{2,3,4,5,6}, Paul I.W. de Bakker^{3,4,7}, Jose C. Florez^{2,3,4,6}, Paul W. Franks⁹, Robert L. Hanson¹⁰, William C. Knowler¹⁰, Zenith Maddipati¹, Toni I. Pollin¹¹, Kathleen A. Jablonski¹

¹The Biostatistics Center, The George Washington University, Rockville, MD; ²Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts; ³Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts; ⁴Department of Medicine, Harvard Medical School, Boston, Massachusetts; ⁵Department of Genetics, Harvard Medical School, Boston, Massachusetts; ⁶Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts. ⁷Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; ⁸Genetic Epidemiology and Clinical Research Group, the Department of Public Health and Clinical Medicine, Division of Medicine, Umeå University Hospital, Umeå, Sweden, and the Department of Clinical Sciences, Lund University, Malmö, Sweden; ⁹Diabetes Epidemiology and Clinical Research Section, National Institute of Diabetes and Digestive and Kidney Diseases, Phoenix, Arizona; ¹⁰Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, Baltimore, Maryland; ¹¹Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, Baltimore, Maryland;
*Contact author: ltipton@bsc.gwu.edu

Keywords: SNP Analysis, Cox Regression, Time-to-event, *Perl*

We improved the performance of a previously developed analytic pipeline to analyze single nucleotide polymorphism (SNP) arrays by converting existing code to *Perl* scripts calling *R* subroutines. A unique feature of our pipeline is that it permits the analysis of time-to-event hypotheses using Cox proportional hazards models with one or more covariates, thus permitting the analysis of SNPs for prospective cohorts and randomized clinical trial study designs.

Several variations of Cox proportional hazards models are implemented in one function in *R* using the **survival** package [1]. One of the covariates may be designated as a primary variable (such as treatment) to perform a variable-by-genotype interaction test. The results are then stratified by this variable. The primary variable can have multiple categories (i.e. more than two treatment groups). The user determines which model is most appropriate for interpretation.

In addition to Cox modeling, PASALS follows a similar format to calculate general linear models for continuous variables and logistic regression for binary variables that may vary by genotype and a primary variable such as treatment group. Both regression models report type III analysis of variance *P* values through use of the **car** package [2]. Taking advantage of the **genetics** package [3], the pipeline also allows for the analysis of allele frequencies by covariates such as sex, treatment and self-reported ethnicity. Results include ethnic-specific tests for Hardy-Weinberg equilibrium and summary statistics on allele and genotype counts.

By switching our code to *R* and *Perl* scripts, performance was greatly improved. Analysis of a test data set with 1,500 SNPs now takes less than 10 minutes where it previously took approximately 12 hours to execute Cox proportional hazards regression to predict an outcome.

This work was funded by RO1 (DK072041-1) "Common Variation in Candidate Genes in the DPP" as a collaboration between Massachusetts General Hospital/Broad Institute, the University of Maryland, NIDDK Phoenix, and The George Washington University.

References

- [1] Terry Therneau (2-02). A Package for Survival Analysis in S. R package version 2.36-12.
- [2] John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>
- [3] Warnes GR. "The Genetics Package," R News, Volume 3, Issue 1, June 2003.

Absolute risk models and calculators for clinical use in *R*

Stephanie A. Kovalchik^{1,*}

1. National Cancer Institute, Rockville, MD

*Contact author: kovalchiksa@mail.nih.gov

In epidemiology, absolute risk models are used to estimate the probability that an individual will develop a disease within a given time interval in the presence of competing events. The development of an absolute risk model from cohort data requires the selection of a relative risk model and the estimation of a baseline hazard function for both the primary and competing event outcomes. The strongest test of the validity of a developed absolute risk model is the evaluation of its calibration and discriminatory ability in an independent cohort. In this talk, I will show methods for developing and validating an absolute risk model in the *R* language, using flexible methodology that can accommodate non-log-linear models of relative risk. The methods are demonstrated with an absolute risk model to predict the risk of second primary thyroid cancer among childhood cancer survivors.

Once an absolute risk model has been validated, it could be a useful tool for clinical decision-making. However, to be practically useful, this tool has to be easy to access and easy to implement. Using the *R* package **fgui** of Hoffmann and Laird [1], I show how to program a user-friendly absolute risk calculator that can be used in clinical settings to obtain predictions and uncertainty intervals for the disease risk of individual patients.

Keywords: absolute risk, survival analysis, GUI

References

- [1] Hoffmann, T. J. and N. M. Laird (2009). **fgui**: A method for automatically creating graphical user interfaces for command-line *R* packages. *Journal of Statistical Software* 30(2), 1–14.

Subscription Survival Analysis in R

Jim Porzak

Senior Data Scientist, Viadeo.com
Contact author: jporzak@gmail.com

Keywords: marketing, survival, subscription, customer.

Survival analysis started in reliability engineering and medical research [1]. More recently it has been used by marketers to better understand customers in subscription based businesses [2, 3]. While the basic math doesn't change, there are practical differences when applied to marketing. First, N is huge. Secondly, there are generally many cohorts driven by marketing variations around product, offer, price, acquisition source, and various subscriber properties. Thirdly, the assumption of hazard stability over time certainly doesn't hold.

In order to demonstrate subscription survival techniques with non-proprietary data, we first needed a tool to generate pseudo-random subscriber data. `subsurvGen` generates data that closely mimics real-world situations based on the choice of generation parameters.

Since N is huge, we use non-parametric methods to calculate hazard and survival. Emphasis is placed on marketer-friendly display of results – both graphically and in tabular form. If subscription average sales price is known, the n-year long term revenue (LTR) is also calculated.

Tactical marketers need survival and LTR at the most granular cohort. Strategic marketers, on the other hand, are interested in LTR at global or regional level. `subsurvRollup` supports flexible roll-ups across multidimensional cohorts to any level.

Hazard space is much more aligned with the marketer's world. Their actions to improve retention are measured by changes in hazard. “What-if” modeling, like “if we reduced churn by 10%,” is supported by `subsurvModel` which gives the projected impact on n-year average tenure and LTR.

There is a challenging dilemma between the strategic marketer's need for a multi-year LTR – which is one limit on allowable acquisition cost – and the tactical need to measure short term changes in hazard. `subsurvModel` also extrapolates short term hazards out to n-years based on the historical decay in hazard space.

The presentation concludes with examples of how subscription survival analysis answered real-world marketing questions.

References

[1] Wikipedia. http://en.wikipedia.org/wiki/Survival_analysis.

[2] Linoff, G, Survival Data Mining for Customer Insight. *Intelligent Enterprise* 7-12, August 2004.

[3] Porzak, J, Subscription Survival Modeling for Fun and Profit. Presented at Predictive Analytics World, San Francisco, March 2012.

HiveR: 2 and 3D Hive Plots of Networks

Bryan A. Hanson*

Dept of Chemistry & Biochemistry, DePauw Univ. Greencastle Indiana USA

*Contact author: hanson@depauw.edu

Keywords: Hive plots, social networks, food webs, bioinformatics

HiveR is an *R* package for creating and plotting 2D and 3D hive plots. Hive plots are a unique method of displaying networks of many types; the concept was developed by [Martin Krzywinski](#) at the Genome Sciences Center. The key innovation in a hive plot is that nodes are deliberately assigned to an axis rather than being positioned by an algorithm whose results depend upon the position of the other nodes. Thus a node is plotted on a particular axis, has a radius along that axis, and may have color and size. Assignment to an axis is based upon the context of the problem, for instance membership in a certain category, while radius, color and size can be a function of network properties, or any other quantitative measure appropriate to the problem. The advantage of this approach is that deletion or addition of a node does not cause the entire layout to change and as a result, comparison of related networks using is straightforward (and implemented in a hive panel). Finally, edges connecting nodes may have a width and a color. The net result is that hive plots are rational and predictable as the layout depends only on the structural properties of the network. They are also very flexible and may be tuned to show particular properties of interest: the mapping of network properties is primarily limited by one's creativity, insight, and the particular knowledge domain. **HiveR** is a fresh *R* implementation of Krzywinski's *Perl* program. It extends the original notion of 2D hive plots to include interactive 3D plots using **rgl**, which can also be made into animations. Examples of 2D and 3D hive plots from a range of fields will be presented. **HiveR** is available on CRAN. In the event you can't attend the talk or wish to chat, there is also a poster on **HiveR**. Stop by and share your ideas, as there are many features that could be added to **HiveR**.

References

M. Krzywinski, I. Birol, S. Jones & M. Marra (2011). Hive Plots - Rational Approach to Visualizing Networks. *Briefings in Bioinformatics* [doi:10.1093/bib/bbr069](https://doi.org/10.1093/bib/bbr069)

RHive in a data scientist's tool box

Seonghak Hong^{1,*}, Heewon Jeon²

1. NexR Corp.

*Contact author: Aiden.hong@nexr.com

Keywords: *R*, *Hive*, *Hadoop*, *R*, **RHive**, Big Data

RHive : *R* and *Hive*

This session introduces the **RHive** project, which integrates *R* with *Hive*. *Hive* can deal with Hadoop data via the *SQL*-styled *HQL*. By expanding *Hive* does **RHive** enable its usage in *R*, allowing analysts to use their familiar *SQL* to process data in *R*. This session not only introduces **RHive**'s structure and features but also a case of actual analysis tutorial by using core **RHive** API.

We are currently facing the need for analyzing massive data by the terabyte. Though we can use well known ETL analysis frameworks for massive data, such as *Hadoop* and *Hive*, but handling them is hardly easy, much less integrating with *R*. Hence analyzers not only have to spend time analyzing data, but also for the aforementioned peripheral tasks.

RHive is an *R* package for connecting *R* and *Hive* together. It makes optimal usage of *R*'s syntax to handle *Hive*, and is also able to minimize the usage of *SQL* syntax to increase *R*'s approachability towards *Hive*. Using **RHive** grants Big Data analyzers an easier time in an *R* environment.

This session will introduce how **RHive** works and use cases.

References

- [1] NexR (2011). RHive : R and Hive, <http://cran.r-project.org/web/packages/RHive/index.html>.
- [2] NexR (2011). RHive tutorial, <https://github.com/nexr/RHive/wiki> .

Providing scalable, transparent access to database-resident data for efficient enterprise advanced analytics

Mark Hornick*, Denis Mukhin, Patrick Aboyoun, Vaishnavi Sashikanth

Oracle

*Contact author: mark.hornick@oracle.com

Keywords: database, transparency, R, scalability, SQL

Data analysts and statisticians, performing advanced analytics on enterprise data, are often affected by reduced efficiency due to three key pain points associated with accessing database-resident data: access latency, paradigm switching, and ability to scale to large datasets in their tool of choice, i.e., *R*.

Access latency occurs when analysts cannot directly access database data, but must request data extracts, typically from IT organizations. This impedes progress until the requested data is delivered. Enterprise costs, sometimes hidden, resulting from access latency include exporting, moving, and storing these data extracts, along with the associated backup, recovery, and security risks.

Paradigm switching involves changing from the *R* language and environment to a *SQL*-based language and/or environment for data access. Here, *R* users specify *SQL* to formulate queries to process or filter data in the database, and then pull data into the *R* environment for further processing using *R*.

When dealing with large enterprise-level data sets, *R* often does not scale, limited by the memory of the machine where *R* executes and not able to leverage multiprocessor machines without special packages or programming. With packages such as **foreach**, **snow**, **Rmpi**, *R* users may need to restructure their scripts or be aware of the underlying hardware configuration to enable parallelism.

These pain points can be minimized, if not eliminated, by providing transparent access to database tables from *R*. In our talk, we explain how *Oracle R Enterprise* [1] provides access from base *R* functionality to Oracle Database tables and views, allowing direct access to database-resident data from *R*. Knowledge of *SQL* or data tools is not required, eliminating paradigm shift. Using the database as a compute engine enables scaling to data sets much larger than available RAM and to leverage multiprocessors for parallel query execution, as well as user-controlled data parallel execution.

Through the *Oracle R Enterprise* transparency layer, *R* users access data stored in the database as virtual data frames, matrices, vectors, etc. Base *R* functions performed on, e.g., `ore.frame` – an S4 class wrapping a database table or view with `data.frame`-like methods, are overloaded to generate *SQL*. This *SQL* is transparently sent to Oracle Database for execution. This transparency is complemented with the ability to embed *R* scripts in Oracle Database for execution using database server-resident *R* engines, which can be spawned in parallel. Overall functionality is expanded to include a wide range of statistical functions and machine learning algorithms executed in Oracle Database from *R* via Oracle *SQL* extensions.

In this talk, we introduce the **ORE** package, and the underlying architecture and high level design for *Oracle R Enterprise* transparency, embedded *R* execution, and enhanced statistical capabilities.

References

[1] Oracle (2012). Oracle R Enterprise, www.oracle.com/technetwork/database/options/advanced-analytics/r-enterprise.

Parallel Apply Functions for In-Database Distributed Scoring

Patrick McCann^{1,2,*}

1. eXelate
2. Virginia Tech

*Contact author: analytics@exelate.com

Keywords: MapReduce, Parallel Computing, Netezza, Apply Functions, High Performance Computing

A common task in predictive analytics is to apply a trained classification or regression model to a set of unlabeled rows to determine a model score, inferred class, or predicted value. In many applications, the unlabeled row count can be extremely large. This type of task is “embarrassingly parallel” in that the operations on groups of unlabeled rows are not dependent on each other and the work can be spread out in a distributed system. In *R*, this task is often handled by the `predict` function in the **stats** package, or a method written for this function within a package or by an analyst. A member of the `apply` family of functions would typically be used to “apply” the prediction function to elements of a data structure. Several packages have implemented parallel versions of some of these functions, for example `emrapply` in the **segue** package and `mclapply` in the **multicore** package.

Using the **nza** package’s `nzApply` & `nzTapply` functions we can push arbitrary *R* functions to a database hosted on a Netezza Twinfin appliance. These functions have similar respective functionality to the `apply` & `tapply` functions in the **base** package. We find two advantages to this framework: minimization of data transfer by moving the function to the data as well as the computation speedups associated with distribution of an embarrassingly parallel problem. In particular, we have used this framework to score user attributes using models trained for the purpose of targeting online advertising.

References

- [1] Cezary Dendek, Przemysław Biecek, Paweł Chudzian, and Justin Lindsey (2010). Massively parallel analytics for large datasets in *R* with **nza** package, <http://www.r-project.org/conferences/useR-2010/slides/Biecek+Chudzian+Dendek+Lindsey.pdf>
- [2] Lukasz Bartnik (2009). *R* on Netezza’s TwinFin Appliance, <http://www.biecek.pl/WZUR2009/LukaszBartnik2009c.pdf>

Introduction to R for Data Mining

Joseph Rickert^{1*}

1. Revolution Analytics

*Contact author: joseph.rickert@revolutionanalytics.com

Keywords: Teaching R, Data Mining, clustering, classification, big data

The R language is often portrayed as having a steep learning curve. Not only does this curb enthusiasm among beginners, it misrepresents the modular structure of the language and the ability to accomplish serious work with relatively modest R language skills. In much the same way that an adult learns a foreign language, by setting achievable goals and getting constructive feedback, anyone with a reasonable goal can make quick process with R. This talk is about teaching R to professionals who want to begin using R as a data mining platform. It will start with the point-and-click Rattle interface and quickly move to writing simple R code to accomplish some serious data mining tasks with classification trees, support vector machines and randomForests. The talk will conclude with building a glm model on a large data file.

Outline :

- Painting the big picture for teaching R
- Point-and-click datamining {rattle}
- Simple R code :
 - Classification trees with `rpart{rpart}`, `drawTreeNodees{rattle}` and `performance{ROCR}`
 - SVM classification with `ksvm{kernlab}` and `train{caret}`
 - Ensemble model with `randomForests{randomForests}`
 - Logistic regression for big data with `rxLogit{revoScaleR}`

References

- [1] Rickert, Joseph (2011). The RevoScaleR Data Step Whitepaper, <http://www.revolutionanalytics.com/why-revolution-r/whitepapers/Data-Step-White-Paper.pdf>
- [2] Williams Graham (2011). Data Mining with Rattle and R, The Art of Excavating Data for Knowledge Discovery. Springer

hdlm: User Oriented, High-Dimensional Linear Model Estimation

Taylor Arnold^{1,*}

1. Yale University, Department of Statistics

*Contact author: taylor.arnold@yale.edu

Keywords: model selection, high performance computing, Bayesian statistics, text mining

High dimensional model selection algorithms, such as the lasso, have garnered much attention in recent years. A large number of *R* packages for conducting such selection algorithms exist, however they unanimously concentrate on the construction of simple point estimators while ignoring useful modifications for data analysis such as scalability, methods for plotting the output, and (most importantly) the production of regression tables. The package **hdlm** [1] rectifies this gap by providing to practitioners: (i) a function to efficiently and easily create sparse regression tables, (ii) an ‘hdlm’ class with over two dozen methods such as plot and predict, and (iii) a generic technique for alternating the default behavior by incorporating new model fitting routines without needing to directly modify the **hdlm** package’s source code.

Functionally, we provide two default behaviors for constructing high dimensional regression tables, both of which were previously unavailable in *R*. The first uses the two-stage approach of Wasserman and Roeder [6], with the generalization proposed by Meinshausen et. al. [4] to increase robustness. The first stage of this method by default uses the elastic net with the **glmnet** package [2], while the second stage uses the standard `lm` function from the package **stats**; both can be easily modified. The second implemented method constructs regression tables using a latent variable Bayesian approach solved via Gibbs Sampling MCMC, as first suggested by Lee et al. in an application to gene expression levels [3].

The talk will focus on design choices made in the package as well as relevant computational issues. As an example of the ability to customize the underlying machinery of the package **hdlm**, we will show how to replace the default elastic net / linear model construction with one based on quantile regression. We also demonstrate how careful use of sparse matrix representations and parallelism via the **foreach** package [7] can be used to greatly increase usability when presented with relatively large datasets. Finally, we end by illustrating a practical application to text mining using the Google N-gram dataset [5].

References

- [1] Arnold, T. B. (2012). *hdlm: Fitting High Dimensional Linear Models*. R package version 1.1.
- [2] Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1), 1.
- [3] Lee, K., N. Sha, E. Dougherty, M. Vannucci, and B. Mallick (2003). Gene selection: a bayesian variable selection approach. *Bioinformatics* 19(1), 90–97.
- [4] Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* 104(488), 1671–1681.
- [5] Michel, J., Y. Shen, A. Aiden, A. Veres, M. Gray, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. (2011). Quantitative analysis of culture using millions of digitized books. *science* 331(6014), 176.
- [6] Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Annals of statistics* 37(5A), 2178.
- [7] Weston, S. (2011). *foreach: Foreach looping construct for R*. R package version 1.3.2.

Rules, Rules, Rules! Regression Modeling Using the Cubist Package

Max Kuhn¹ and Chris Keefer²

1. Nonclinical Statistics, Pfizer Global R&D
 2. Pharmacokinetics, Dynamics and Metabolism, Pfizer Global R&D
- *Contact author: max.kuhn@Pfizer.com

Keywords: Machine Learning, Regression, Model Trees

Machine learning (aka pattern recognition)(aka predictive modeling) has a long history with tree-based models, such as CART or random forests. However, a parallel set of methodologies have been developed that use rules to create effective predictions (e.g. Quinlan (1992) and Quinlan (1993)). Many of these models are not well known and have not undergone significant peer review but are nonetheless very powerful. Recent developments are bringing a revival of these models. This talk will discuss rule-based models using a port of the **Cubist** (and **C5.0**) code to *R*.

References

- Quinlan, J. (1992). Learning with continuous classes. *Proceedings of the 5th Australian Joint Conference On Artificial Intelligence*, 236–243.
- Quinlan, J. (1993). Combining instance-based and model-based learning. *Proceedings of the Tenth International Conference on Machine Learning*, 236–243.

Finding the number of clusters in a data set : **Nb.clusters** package

Malika Charrad^{1,*}, Nadia Ghazzali¹, Veronique Boiteau¹, Azam Niknafs¹

1. Universit Laval

Department of Mathematics and Statistics[-2pt]

*Contact author: malika.charrad.1@ulaval.ca

Keywords: Number of clusters, Validity Indices, Cluster validity, Kmeans, Hierarchical clustering.

Clustering is the partitioning of a set of objects into groups (clusters) so that objects within a group are more similar to each others than objects in different groups.

Most of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set. As a consequence, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity. In general terms, there are three approaches to investigate cluster validity. The first is based on external criteria, which consist in comparing the results of cluster analysis to externally known results, such as externally provided class labels. The second approach is based on internal criteria which use the information obtained from within the clustering process to evaluate how well the results of cluster analysis fit the data without reference to external information. The third approach of clustering validity is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameters values, e.g. the number of clusters.

In the literature, a wide variety of indices have been proposed to find the optimal number of clusters in a partitioning of a data set during the clustering process. Although a vast number of references exist, few comparative studies have been performed on these indices [5]. Moreover, for most of indices proposed in the literature, programs are unavailable to test these indices and compare them.

The R package, **Nb.clusters**, has been developed specifically for that purpose. It implements 30 indices for cluster validation ready to apply on outputs produced by clustering algorithms, Hierarchical clustering and Kmeans, coming from the same package. Most of these indices are described in Milligan and Cooper study [5]. The `Nb.clusters` function allows to apply one or 30 indices simultaneously and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures ("euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski"), and clustering methods ("ward", "single", "complete", "average", "mcquitty", "median", "centroid").

References

- [1] Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal Cybern*, 95–104.
- [2] Halkidi, M., I. Batistakis, and M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* 17(2/3), 107–145.
- [3] Kaufman, L. and P. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. New York, NY, USA: Wiley.
- [4] Lebart, L., A. Morineau, and M. Piron (2000). *Statistique exploratoire multidimensionnelle*. Paris, France: Dunod.
- [5] Milligan, G. and M. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2), 159–179.
- [6] Tibshirani, R., G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2), 411–423.

Power and Sample Size for Safety Registries: new methods using confidence intervals and saw-tooth power curves

Paul Schuette^{1*}, C. George Rochester¹, Matthew Jackson¹

1. Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US FDA
*Contact author: Paul.Schuette@fda.hhs.gov

Keywords: Power, Sample Size, Post-market Safety

In 2002 Chernick and Liu, [1] showed that power functions for discrete data can exhibit nonmonotone saw-toothed behavior. We expand on Chernick and Liu's work, presenting new methods which take advantage of the duality between hypothesis tests and confidence intervals. Specifically, we present a new procedure for generating a power curve for a test of a parameter for a discrete distribution. This procedure is applicable to any test method, so long as that test is equivalent to some confidence interval method under the aforementioned duality.

We illustrate our results for a binomial parameter using Clopper-Pearson, Agresti-Coull, Wilson and Blaker intervals. Algorithms for calculating Blaker confidence intervals appear to be only available in *R*, through the packages **BlakerCI**, **exactci** and **PropCIs**. We present similar results for Poisson rates, using exact Garwood intervals and asymptotic score intervals. Additionally, we discuss how the saw-toothed power functions we obtain may be used in a regulatory context for power and sample size determination of post-market product safety registries and provide recommendations in the rare event setting. Our results in this area modify and extend the work of previous authors such as [2]. We also present findings regarding the relationship between detectable effect size and sample size for a given level of power. Our results are implemented in *R* and validated with *MATLAB*.

The authors acknowledge the support of the Office of Women's Health at the US FDA.

References

- [1] Chernick, M. R. and C. Y. Liu (2002). The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportions using exact methods. *The American Statistician* 56, 149–155.
- [2] Moride, P. T.-B. B. B. Y. and L. Abenhaim (1994). Sample size calculations for single group post-marketing cohort studies. *J. Clin Epidemiol* 47, 435–439.

autoLDA, an *R*-package for Automated Longitudinal Data Analysis

Shubing Wang*, Andy Liaw

Biometrics Research Department, Merck Research Laboratories (MRL)

*Contact author: shubing.wang@merck.com

Keywords: Longitudinal data analysis, mixed-effects modeling, baseline adjustment, multiplicity adjustment

Longitudinal studies are among the most widely used in the pharmaceutical industry and are absolutely essential for making decisions on drug efficacy and safety. They are complex, diverse and often nuanced in their design. Proper statistical analysis of these studies therefore requires high level statistical expertise and can be time consuming. In order to structurally shorten the processing time and to make this expertise readily available for bench scientists, we developed an *R*-package: **autoLDA**, which implements the most widely-used statistical methodology for longitudinal data analysis, including mixed-effects modeling and multiple hypothesis testing. The main challenge in achieving this goal is the transfer of statistical expertise in a flexible way that does not compromise the scientific quality of analysis. The main body of the package is a graphics user interface (GUI) that is very intuitive and user-friendly which includes various options to accommodate different study designs, data formats, and reports. The package also provides various visualization tools to explore the data in order to reveal crucial information and abnormalities. It allows users to choose between longitudinal models with different baseline adjustments, within-group covariance structures, measurement scales and output formats. Established as well as the novel methods based on False Discovery Rate, Bootstrapping and Random Field theory, are used for multiplicity adjustment. Finally, this package automatically generates PowerPoint slides with analysis report, including plots, tables and summary of findings. The software has been extensively used for analysis of longitudinal studies with the data obtained by ultrasound, telemetry, micro-CT, and PET imaging modalities across Merck Research Laboratories. It was found to reduce duty cycle from the order of weeks to a couple of days, thus enabling research teams to deliver results and drive decisions at a much faster pace. Results of these analyses were used by numerous programs of BRIE, CVD, and Neuroscience franchises. Examples of the analyses are presented and discussed.

Developing a context-specific plug-in for R Commander: Lessons learned from a longitudinal data analysis training in a resource-limited setting

Sandra D Griffith^{1*}, Sarah J Ratcliffe¹

1. Department of Biostatistics & Epidemiology, University of Pennsylvania

*Contact author: sgrif@upenn.edu

Keywords: Teaching, R Commander, Global Health

Capacity-building global health partnerships often involve short-term training visits to provide scientific investigators with the quantitative skills necessary for data analysis. In choosing software for such trainings, one must consider several important constraints: the burden of license costs in resource-limited settings; the need to cater to investigators with varying programming backgrounds; and the desire for flexibility as course content often changes in real time. *R*, coupled with the **Rcmdr** package, can overcome these obstacles.

We discuss these issues in the context of a recent training visit with scientific investigators in the Botswana-UPenn partnership. The week-long training aimed to provide investigators with the tools to analyze longitudinal data on behavioral interventions for HIV prevention in Botswana. Most investigators had very limited or no exposure to *R* or other command-based software, and felt most comfortable using the point-and-click functionality of *SPSS*. This motivated our use of *R* and the **Rcmdr** package which provide a freely available GUI familiar to *SPSS* users. The longitudinal nature of the data, however, required *R* functionality unavailable in **Rcmdr**. Although the flexibility of **Rcmdr** permits development of customized plug-in packages to extend its functionality, none existed for our particular needs.

Within a limited time frame and continually evolving requirements, we developed a plug-in package with only the functionality necessary for this specific training. Hosting the package in a local CRAN repository allowed participants to easily install the plug-in, as well as obtain daily updates as necessary. We discuss the successes and challenges faced using this approach and the implications for flexible software development and distribution strategies for similar trainings in the future. Although the software developed for this training was highly context-specific, the experience motivated its extension to **RcmdrPlugin.Longitudinal**, a full package for longitudinal data analysis using **Rcmdr**.

References

Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. *Journal of Statistical Software*, 14(9): 1–42.

EZ-R4Excel: unleashing the statistical power of R in Excel

Jyotsna Kasturi^{*}, Davit Sargsyan, Mariusz Lubomirski, Dhammika Amaratunga, Bill Pikounis

Non-Clinical Statistics, Janssen Pharmaceutical companies of Johnson & Johnson, United States

*Contact author: jkasturi@its.jnj.com

Keywords: Excel, R, VBA, custom statistical analyses

Microsoft Excel offers a user-friendly, convenient and flexible spreadsheet environment and an industry standard for data storage, management and analysis. It is widely used by the pharmaceutical industry in a variety of regulatory and non-regulatory applications.

EZ-R4Excel is a novel and user-friendly software especially applicable to non-regulatory use. It offers the unparalleled capability of linking an Excel Spreadsheet with the statistical package R, thus enabling a tremendous amount of flexibility in statistical analysis to be performed directly inside the Excel environment. The application is developed as a simple Excel add-in. Its implementation is an extension of custom-written VBA macros seamlessly integrated with R scripts in the background. EZ-R4Excel is a powerful engine significantly enhancing the statistical power of Excel within its already familiar user interface.

The conceptual design of EZ-R4Excel is an easy-to-use extensible scaffolding framework, allowing new statistical methods to be easily incorporated into the available set of tools. The vast library of publicly available R packages may be leveraged in addition to offering the capability of integrating novel and customized statistical methods, thus making it potentially applicable for open source licensing. Another design feature of the software is the accessibility of calling functions either by means of GUI driven menus or directly from spreadsheet cells similar to running standard Excel functions, making it very attractive to both expert and novice users.

EZ-R4Excel has demonstrated particular relevance in drug discovery applications to perform specialized analysis methodologies not usually available in Excel such as in-vivo pharmacology, high throughput screening and visualization, etc. The base version of EZ-R4Excel also provides various methods for exploratory analyses and advanced graphical features.

A Simple RExcel Interface for Bayesian Adaptive Clinical Trial Enrollment

Jocelyn Sendeck^{1,2,*}, Terry Hyslop¹

1. Thomas Jefferson University, Division of Biostatistics, Philadelphia PA

2. Temple University, Philadelphia PA

*Contact author: jocelyn.andrel@kimmelcancercenter.org

Keywords: Clinical Trial, Adaptive Bayes Randomization, RExcel

Traditional clinical trial randomization typically involves creating a list of subject assignments to treatment arm based on simple, stratified, or block designs, then assigning each subject accordingly as s/he enters the study. Bayesian Adaptive randomization is rapidly becoming more popular because of its ability to implement early stopping rules through the incorporation of ongoing study outcomes into a prior probability. In practice, having to calculate such a prior eliminates the possibility of easy-to-use randomization envelopes, so we look to a combination of *R* and Excel (RExcel) to make subject enrollment using adaptive randomization simple for the clinical trial coordinator. We have created an Excel workbook that incorporates a set of initial “straight randomization” subject assignments and their study outcomes followed by adaptive subject randomization. Assignment is performed by a customized *R* package that builds off of packages **blockrand** and **zelig**. There may be existing software packages that already do what we have described here. However, we believe that the ubiquity of Excel makes is a convenient and recognizable interface for clinical trial enrollment, especially combined with the power of *R* behind it.

References

Heiberger, R.M. and Neuwirth, E. (2009). *R through Excel: A spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer.

Visualization of Regression Models Using visreg

Patrick Breheny^{1,*}, Woodrow Burchett¹

1. Department of Statistics, University of Kentucky

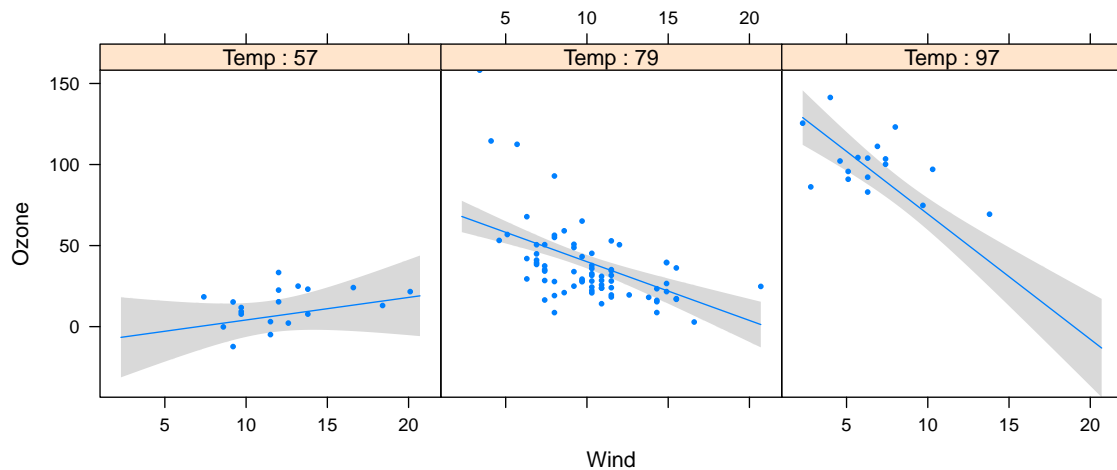
*Contact author: patrick.breheny@uky.edu

Keywords: Graphics, visualization, linear regression, generalized linear models

Regression models allow us to isolate the relationship between the outcome and an explanatory variable while the other variables are held constant. Here, we introduce an *R* package, **visreg**, for the convenient visualization of this relationship via short, simple function calls. In addition to estimates of this relationship, the package also provides pointwise confidence bands and partial residuals to allow assessment of variability, outliers, and deviations from modeling assumptions. The package also provides several options for visualizing models with interactions, including lattice plots, contour plots, and both static and interactive perspective plots. The implementation of the package is designed to be as generic as possible, allowing visualization not only of linear models, but of generalized linear models (*glm*), proportional hazards models (*coxph*), generalized additive models (*gam*), robust regression models (*r1m*), and more.

To provide a quick example:

```
> fit <- lm(Ozone ~ Solar.R + Wind*Temp, airquality)
> visreg(fit, "Wind", by="Temp")
```



Some Challenges of Using *R* in a Regulatory Environment

Jae Brodsky

Office of Biostatistics, Food and Drug Administration
jae.brodsky@fda.hhs.gov

Keywords: FDA, R, SAS, regulatory agency, government

Food and Drug Administration (FDA) statisticians and reviewers currently use several different statistical programs, including *SAS* and *R*. Although *R* meets the qualification, validation, and verification requirements in the Code of Federal Regulations (CFR) Title 21, part 11 [1], there are several additional issues that affect the use of *R* in a regulatory agency.

R and *SAS* each have advantages in different areas of statistical analysis, and statistical work at the FDA could benefit from *R* and *SAS* being used together. For example, *SAS* could be used to perform analytic functions on large data sets that are commonly received from sponsors while *R* could be used for complementary graphical analyses. Impediments to this type of analytic set-up include issues with certifying non-base *R* and selected core packages [2], differences in statistical functions between *R* and *SAS*, the use of legacy code at the FDA, and several different issues due to the fact that *R* is freeware. Additionally, the FDA has historically used *SAS* and there are many reviewers who have never used *R*.

I will discuss my personal experiences as an *R* user at the FDA, more regulatory issues with *R* beyond what is covered in 21 CFR 11, and discuss some possible future approaches to integrating the use of *R* and *SAS* at the FDA.

This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

References

Code of Federal Regulations (2011). 21 CFR Part 11, <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=11>.

The R Foundation for Statistical Computing (2008). R: Regulatory Compliance and Validation Issues - A Guidance Document for the Use of R in Regulated Clinical Trial Environments, <http://www.r-project.org/doc/R-FDA.pdf>.

Econometrics of Network

TszKin Julian Chan

1. Department of Economics, Boston University

✉ Contact author: ctszkin@gmail.com

Keywords: Econometrics, Peer Effect, Network Data

Recently, several econometric methods are proposed to estimate peer effect model with social network data. We include some of them in the package. Also a strategic network formation model is included.

The peer effect model studies how the outcome variable affected by characteristics of peers. Manski [6] separate peer effects into contextual effect (i.e. the influence of exogenous peer characteristics), endogenous effect (the influence of peer outcomes) and correlated effect (factors affect individuals in the same reference group). Identification problems arise in the linear-in-mean model. Lee [5], Graham [4], Bramoullé et al. [1] propose using group size, variance structure and social network structure to identify model respectively.

The network formation model study what factors determine whether two individuals are connected. In the package, we implement a strategic network formation model proposed by Christakis et al. [3]. In this model, agents meet sequentially and decide whether to connect in each period. We can estimate the preference parameters in the model with the MCMC methods.

In 2011, Chan and Lam [2] constructed a dataset with social network structure, academic outcomes and other characteristics of about 1000 high school students in Hong Kong. Using the dataset, we applied the above methods to study the peer effect in academic outcomes.

References

- [1] Bramoullé, Y., H. Djebbari, and B. Fortin (2009, May). Identification of peer effects through social networks. *Journal of Econometrics* 150(1), 41–55.
- [2] Chan, T. K. J. and C. S. T. Lam (2011). Estimating peer effects on academic performance through an endogenous network: Evidence from a student network survey in Hong Kong.
- [3] Christakis, N. A., J. H. Fowler, G. W. Imbens, and K. Kalyanaraman (2010). An Empirical Model for Strategic Network Formation. *Working Paper* (May).
- [4] Graham, B. S. (2008). Identifying Social Interactions Through Conditional Variance Restrictions. *Econometrica* 76(3), 643–660.
- [5] Lee, L. (2007, October). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics* 140(2), 333–374.
- [6] Manski, C. F. (1993). Identification of Endogenous Social Effects : The Reflection Problem. *The Review of Economic Studies* 60(3), 531–542.

Towards an R-based data operating system for the the science clouds

Karim |Chine¹

1. Cloud Era Ltd

*Contact author: karim.chine@gmail.com

Keywords: Cloud Computing, GUIs, e-Science, AaaS

The cloud means affordable and ubiquitous access to “infinite” compute and storage resources. It also means that hardware becomes like software and that infrastructure becomes programmable/scriptable.

R is at the same time a universal glue for software components/libraries and a powerful scripting language that could be used to pilot the cloud and to build programmatically virtual infrastructures for data processing. *R* can also be used to create and assemble on the fly distributed software components and the *R* platform can become a highly productive Data-centric applications and services factory.

The **elasticR** package combined with the Elastic-R portal proposes to take *R* to this new realm. The new *R* package makes it possible to use the Amazon cloud programmatically from a regular *R* session. *R* servers with a rich and stateful interface can be created on *EC2* with simple *R* functions and used to offload time-consuming computations to machines of large capacities, to apply *R* functions to large data sets in parallel and to collaborate in real-time. More interestingly, the elasticR package can be used to build and publish cloud-based applications and services on the fly.

The presentation will give an overview of the package and will discuss new directions for using *R* in the context of e-Science.

References

- [1] Karim Chine (2010). Elastic-R Platform, <http://www.elastic-r.net>.
- [2] Karim Chine (2010). Open science in the cloud: towards a universal platform for scientific and statistical computing. In: Furht B, Escalante A (eds) Handbook of cloud computing, Springer, USA, pp 453–474. ISBN 978-1-4419-6524-0.
- [3] Karim Chine (2010). Learning math and statistics on the cloud, towards an EC2-based Google Docs-like portal for teaching / learning collaboratively with *R* and Scilab, icalt, pp.752-753, 2010 10th IEEE International Conference on Advanced Learning Technologies.

Profile likelihoods and the likelihood paradigm: The *R* Package **ProfileLikelihood**

Leena Choi^{1,*}

1. Department of Biostatistics, Vanderbilt University

*Contact author: leena.choi@vanderbilt.edu

Keywords: nuisance parameter, confidence interval, likelihood support interval, **ProfileLikelihood**

The standard procedure for statistical inference, such as the Wald test or its corresponding confidence intervals (CIs), which is based on the asymptotic normality of the maximum likelihood estimate (MLE), may behave very poorly in certain situations, typically with small samples or when the log likelihood function is highly skewed (i.e., non-normal). When the normality fails, many authors have advocated the direct use of the likelihood for making inference. When the likelihood function is indexed with a single parameter, it is straightforward to make likelihood-based inference. With most practical problems, however, the models often have a vector of parameters that can be partitioned as a small subset of the parameters of interest and nuisance parameters; these nuisance parameters need to be eliminated. A profile likelihood is a promising method for eliminating nuisance parameters, since it asymptotically shares good frequency properties of the likelihood. The *R* Package **ProfileLikelihood** [1] provides profile likelihoods for a parameter of interest in commonly used statistical models, including linear models, generalized linear models, proportional odds models, linear mixed-effects models, and linear models for longitudinal responses fitted by generalized least squares. We will present how to make likelihood-based inference using profile likelihoods and their CIs obtained from the *R* Package **ProfileLikelihood**. We will also illustrate how to measure statistical evidence using standardized profile likelihoods and the k th likelihood support intervals with the likelihood paradigm advocated by Royall [2].

References

- [1] Choi, L. (2011). *ProfileLikelihood: Profile Likelihood for a Parameter in Commonly Used Statistical Models*. R package version 1.1.
- [2] Royall, R. M. (1997). *Statistical evidence: a likelihood paradigm*. Chapman & Hall/CRC.

The SAScii Function

Anthony Damico^{1,*}

1. Kaiser Family Foundation

*Contact author: adamico@kff.org

Topic Area: Statistics in the Social and Political Sciences

Keywords: Importation, US Government Data, SAS, ASCII, Fixed Width Formatted Data

Research Objective: When an ASCII-formatted data set includes only SAS import instructions, an analyst desiring to import the data into R must either have access to the proprietary SAS software or undergo the cumbersome process of creating the parameters of the **foreign** package's `read.fwf` function by hand. R users working at organizations with SAS licenses might read in the data with SAS, then export it immediately as a **foreign** package-compatible XPT or CSV file. Analysts without access to SAS software might be forced to painstakingly convert the column names, widths, and formats from the block of PROC IMPORT code from the SAS script into the appropriate vectors for a `read.fwf` function call. The `SAScii` function eliminates the choice between purchasing expensive software and manually re-formatting SAS import instructions. The objective of this presentation will be to share a brief but useful algorithm that requires two parameters –

- a) a fixed-width format data file
- b) a SAS import script file containing PROC IMPORT instructions

– and returns a fully-formatted data frame inside R.

Implications for R Users: In the spirit of my previous work¹ on the subject of equipping analysts with the tools needed to avoid proprietary software, the `SAScii` function provides a shortcut in the conversion of ASCII-stored data with accompanying SAS import instructions. U.S. government agencies, such as the Centers for Disease Control & Prevention, still release data sets² in only fixed-width format and with only SAS, SPSS, and Stata import statements. The `SAScii` function eases R users' access to data sets lacking importation methods otherwise accessible by open source software.

¹ Anthony Damico. Transitioning to R: Replicating SAS, Stata, and SUDAAN Analysis Techniques in Health Policy Data. *The R Journal*, 1(2):37-44, December 2009.

² http://www.cdc.gov/nchs/nhis/nhis_2010_data_release.htm

Implementation of ANOVA-PCA in R for Multivariate Data Exploration

Matthew J. Keinsley & Bryan A. Hanson*

Dept of Chemistry & Biochemistry, DePauw Univ. Greencastle Indiana USA
*Contact author: hanson@depauw.edu

Keywords: multivariate, PCA, ANOVA, chemometrics, spectroscopy

Both ANOVA and PCA are time-honored, extensively used methods. Recently, Harrington *et. al.* described a method of blending these two techniques for multivariate data exploration. A typical application would be spectroscopic investigation of samples derived from different experimental treatments. The data is partitioned into submatrices corresponding to each experimental treatment in a manner reminiscent of ANOVA; subtraction of these sub matrices from the original data gives a matrix of residual error. The submatrices are then separately subjected to PCA after adding back the residual error. If the effect of a treatment is large compared to the residual error, separation along the 1st PC in the score plot should be evident. With this method, the significance of a treatment can be visually determined. Thus unlike PCA, ANOVA-PCA is not blind to group membership. We implemented the ANOVA-PCA concept using a series of functions in the chemometrics package **ChemoSpec** (available on CRAN). Data is stored in a `Spectra` object. The function `aovPCA` carries out the key matrix manipulations, while `aovPCAscores` and `aovPCAloadings` plot the results. We will demonstrate the method on both simulated and real data sets.

References

Harrington, Vieira, Espinoza, Nien, Romero, and Yergey. "Analysis of Variance–Principal Component Analysis..." *Analytica Chimica Acta* 544.1-2 (2005): 118-27.

Acknowledgements

We are grateful to the Science Research Fellows program and the Chemistry & Biochemistry Department at DePauw for support.

HiveR: 2 and 3D Hive Plots of Networks

Bryan A. Hanson*

Dept of Chemistry & Biochemistry, DePauw Univ. Greencastle Indiana USA

*Contact author: hanson@depauw.edu

Keywords: Hive plots, social networks, food webs, bioinformatics

HiveR is an *R* package for creating and plotting 2D and 3D hive plots. Hive plots are a unique method of displaying networks of many types; the concept was developed by [Martin Krzywinski](#) at the Genome Sciences Center. The key innovation in a hive plot is that nodes are deliberately assigned to an axis rather than being positioned based upon some algorithm. Thus a node is plotted on a particular axis, has a radius along that axis, and may have color and size. Assignment to an axis is based upon the context of the problem, for instance membership in a certain category, while radius, color and size can be a function of network properties. The advantage to this approach is that deletion or addition of a node does not cause the entire network graph to change and as a result, comparison of related networks using hive panels is straightforward. Finally, edges may have a width and a color. The net result is that hive plots are rational and predictable as the layout depends only on the structural properties of the network. They are also very flexible and may be tuned to show particular properties of interest: the mapping of network properties is primarily limited by one's creativity and the particular knowledge domain. **HiveR** is a fresh implementation of Krzywinski's *Perl* program `linnet` using *R*. It extends the original notion of 2D hive plots to include interactive 3D plots using `rgl`, which can also be made into animations. Examples of 2D and 3D hive plots from a wide range of fields will be presented. **HiveR** is available on CRAN.

References

M. Krzywinski, I. Birol, S. Jones & M. Marra (2011). Hive Plots - Rational Approach to Visualizing Networks. *Briefings in Bioinformatics* [doi:10.1093/bib/bbr069](https://doi.org/10.1093/bib/bbr069)

The most general GUI-solution for R

Sebastian Hoffmeister¹

1. Statcon

*Contact author: sebastian.hoffmeister@statcon.de

Keywords: GUI, Usability

R gets more and more attention. Once *R* has been a program for statistician specialists. The greatest part of the community came from the academic field. With the growing acceptance of *R* in the industrial field a new kind of user enriches the community. The typical user from a R&D-division is an expert in chemics, physics or some other natural science. She has some experience in using statistics for her work and knows one of the commonly used statistical packages like **SPSS**, **JMP** or **EViews**.

Usually these potential new *R*-users are looking for a special statistical method, which is not implemented for their software, when they come across *R*. Sadly we loose many of them when they realize that there is only the command line interface. Most of these new users don't want to leave the familiar lane of GUI driven programs like, e.g. **JMP**.

Taking a look at GUI-solutions for *R* the new user finds mainly two kinds of programs. Extended text editors like **RStudio**, **Tinn-R** or **JGR** have a different target audience. These programs are IDEs for *R*-programmers. They do not really want to ease the first steps with *R*. Real GUI-solutions like **RCommander** face the general problem of all menu-driven systems. They are able to support only a limited set of functions. Typically the new user is not interested in *standard* but in very specific functions.

A possible solution for the last drawback will be presented: a concept for a fully-automatic system of menus and user-dialogs. A first implementation will be able to show some of the potential of this idea. Of course this automatic system will never be as comfortable as **R Commander**. Nevertheless ideas to improve the usability of this system and *R* in general will be discussed. The perhaps most impressing refinement is a preview-window. This shows the results of a function while the user is entering the arguments in a user-dialog.

SNA with R Text Mining

Heewon Jeon^{1,*}

1. NexR Corporation

Contact author: madjakarta@gmail.com

Keywords: SNA, Text Mining, CJK

SNA using text mining is the latest trend in Korean society. Used in analyzing twitter data or quotes from talk shows and articles, SNA is used by many Koreans to get a grasp on politicians' true political colors or their inner intentions. For such analyses, the process of extracting keywords and words co-occurring with them and performing calculations for their relationships is an essentiality. Thus a text mining package of a particular language was required: this is the developmental background for the **KoNLP** package. Using this package and a few other existing packages, we've managed to execute a successful text mining SNA, and we are going to provide an overall introduction on them.

This talk will briefly introduce **KoNLP**, **igraph**, and some expectable problems when conducting *R* text mining, using CJK languages (mainly Korean).

1. Safe encoding for CJK and the means we can use system functions for safe use.
2. Introduces **tm**, **KoNLP** and **igraph**.
3. Text mining SNA example using quotes from Korean politicians.

References

- [1] Heewon Jeon (2012). KoNLP Github, <https://github.com/haven-jeon/KoNLP>.

Modeling of Volcanic Sound Radiation Using Finite-difference Time-domain Method in *R*

Keehoon Kim^{1*}, Jonathan M. Lees¹

1. University of North Carolina, Chapel Hill

*Contact author: keehoon@live.unc.edu

Keywords: FDTD, Volcanic Explosion, Infrasound

Finite-difference Time-domain (FDTD) modeling is a powerful technique to numerically solve partial differential equations. The FDTD method is used in a variety of physical applications such as electrodynamics, seismology, and acoustics in order to simulate natural phenomena. We have developed FDTD codes in *R* and applied them to modeling volcanic sound radiation generated by explosive eruptions. Radiation patterns of volcano infrasound may be important for hazard reduction in dangerous regions where volcanic catastrophes are common (e.g. Ecuador, Guatemala, Iceland). The FDTD method involves designing a model domain that includes complex volcano topography and estimation of physical properties across the domain nodes. Our *R* codes provide a simple platform for setting up the equations and executing these complex calculations with a simple interface. Following solution in space and time, post-processing is required to convert numeric results into a form having physical meaning, such as the visualization of the time evolution of the sound pressures and related radiation patterns. The pre- and post-processing are combined seamlessly in our FDTD approach providing a convenient solution to an otherwise complicated FDTD problem. We present the sound modeling results from our *R* FDTD codes with an application of eruption sounds [1] at the very active Karymsky volcano, Russia.

References

- [1] Kim, K. and J. M. Lees (2011, March). Finite-difference time-domain modeling of transient infrasonic wavefields excited by volcanic explosions. *Geophysical Research Letters* 38(6), L06804.

A JVM-based Compiler Strategy for the R Language

Helena Kotthaus^{*}, Sascha Plazar, Peter Marwedel

Computer Science 12, TU Dortmund University
^{*}Contact author: helena.kotthaus@tu-dortmund.de

Keywords: R Language Optimization, Java, Compiler

The *R* programming language has become invaluable for analysis and evaluation of statistical methods. *R* is a multi-paradigm language with functional characteristics, a dynamic type system and different object systems. These characteristics support the development of statistical algorithms and analyses at a high-level of abstraction. Like for many dynamic languages, *R* programs are processed by an interpreter. Especially in the domain of statistical learning algorithms and bioinformatics, e.g. when analyzing high-dimensional genomic data, this interpretation often leads to an unacceptably slow execution of computation-intensive *R* programs. Our goal is to optimize the execution runtime of such *R* programs. Therefore, we plan to develop a JVM-based compiler strategy including an *R*-interpreter written in *Java* and an extensible just-in-time (JiT) compiler.

With the use of an *R*-interpreter written in *Java* [1], the execution of *R* programs could be optimized: By targeting the JVM, JiT compilation is enabled within the interpreter code. However, transferring the *R* interpretation process to the JVM does not automatically lead to high optimization potential, because *R* programs still need to be interpreted. Additional source level optimizations should be applied to the intermediate representation (IR) produced by the *R*-parser. Although JiT compilation could already speed up the interpretation process, the native JiT-compiler is not aware of the *R* language and its specific optimization needs. In order to push more aggressive optimizations, the JiT-compiler should be extended by knowledge about *R* characteristics to enable language specific low-level optimizations and generate highly optimized machine code. For this purpose, the Graal JiT-compiler [2], which is especially designed for extensibility, should be employed.

On our poster we present our optimization ideas and development plans for the JVM-based compiler strategy for the *R* Language.

References

- [1] Bertram, A. (2012). JVM-based Interpreter for the R Language for Statistical Computing. <http://code.google.com/p/renjin>.
- [2] Würthinger, T. (2011). Extending the graal compiler to optimize libraries. In *Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion*, pp. 41–42.

Open Platform for Quality Methodologies: a Six Sigma Framework for Competitiveness

Emilio L. Cano^{1,2,*}, Javier M. Moguerza¹, Andres Redchuk¹

1. University Rey Juan Carlos

2. University of Castilla-La Mancha

*Contact author: emilio.lopez@urjc.es

Keywords: Six Sigma, Quality, Process Improvement, Competitiveness

The competitiveness of Industry is a great challenge in the globalization environment we are living in. The implementation of systematic methods of process improvement is undoubtedly a way to improve the competitiveness within an individual company. These methods are usually gathered into different "Quality Methodologies", and Six Sigma is one of them. Six Sigma has become one of the most successful methodologies for process improvement. ISO standards [3, 4] have recently been published, and a number of publications on this topic have appeared in the last decade. The essence of Six Sigma is the application of the Scientific Method to process improvement and the key to its success is that Six Sigma translates the scientific terminology into a simple way to apply science to process improvement through the DMAIC strategy (Define, Measure, Analyze, Improve and Control). Six Sigma needs statistical software for data analysis and *R* can be the ideal choice for many reasons. In addition to be a stable and reliable statistical software and programming language, it is Open Source. Though some sectors are yet reluctant to adopt Open Source technologies, the fact is that it is being increasingly used in many activity sectors, going beyond its original environment: academia and research. With this aim, and after publishing a book on how to tackle Six Sigma project with *R* [2] and the *R* package **SixSigma** [1], we present in this work a project to develop a complete framework for applying Six Sigma using Open Technologies, where *R* is the main protagonist. The project is focused in supporting specially, but not limited to, SMEs (Small and Medium Size Companies). Simulation and efficiency are crossing topics that are taken into account throughout the project development. The result of the project will be a platform that can be used independently within an organization, or exploited as SaaS (Software as a Service) by third parties. Along with the improvements achieved applying Six Sigma, the open nature of the platform allows to save money in software licences. All in all, enhancing competitiveness.

References

- [1] Cano, E. L., J. M. Moguerza, and A. Redchuk (2011). *SixSigma: Six Sigma Tools for Quality Improvement*. R package version 0.6.0.
- [2] Cano, E. L., J. M. Moguerza, and A. Redchuk (2012). *Six Sigma with R. Statistical Engineering for Process Improvement*, Volume 36 of *Use R!* New York: Springer.
- [3] ISO (2011a). Iso 13053-1:2011 - quantitative methods in process improvement – six sigma – part 1: Dmaic methodology.
- [4] ISO (2011b). Iso 13053-2:2011 - quantitative methods in process improvement – six sigma – part 2: Tools and techniques.

Fitting Sinusoidal Hysteresis in R: The R package **fitellipse**

Spencer Maynes^{*1}, Fan Yang¹, Anne Parkhurst¹

1. Department of Statistics, University of Nebraska-Lincoln

*Contact author: smaynes89@gmail.com

Keywords: Ellipse fitting, Hysteresis, Pattern Recognition, Calibration

An R package **fitellipse** was created to model a sinusoidal hysteretic process. This package implements functions for fitting ellipses based on linear least squares, ellipse-specific non-linear least squares, and two-stage simple harmonic least squares. Ellipse area, along with the hysteretic properties of retention, coercion, and lag can be estimated from these ellipses along with their confidence intervals. A system displays hysteresis if it is influenced by its past and not just its current state. This often results in the formation of a loop when the input variable is periodic in nature. Fitting elliptical loops is useful in many situations where output is bivalued for a given input value; examples include unemployment in response to GDP, heat stress in animals, and the Stirling cycle. Information on how to use **fitellipse** will be provided.

Project Management Factor Text Analysis with R

Rodger A. Oren¹, Marilyn B. Harris²

1. Bureau of TennCare

2. Capella University

*Contact author: roren@mindspring.com

Keywords: Project Management, Earned Value, Earned Schedule, Social Science Research

Project management as a profession has seen considerable growth and interest in the latter half of the 20th century into our present period in the 21st century as organizations seek to release products and services in a predictable manner. The field provides practitioners with skills in estimating, monitoring and controlling the budget, duration and deliverables in the temporary endeavor which is called a project.

Those of a more analytical background seek out metrics, such as Earned Value and Earned Schedule, to help monitor and forecast the direction of cost and duration parameters. Kerzner [1] notes that Earned Value has proven itself to work in complex projects in fields such as construction and software, providing monetary values which are useful for cost and schedule considerations. An emergent concept, Earned Schedule from Lipke [2], seeks to utilize a similar algebraic ratio approach to calculate duration, solving an inherent shortcoming of the temporal component of Earned Value.

Success is more elusive than failures in the field. Practitioners grapple with how to increase the former while reducing the latter results without any way of knowing what is really contributory to success versus just window dressing. Due to this situation and emerging interest in the academic setting, the project management field is in its beginning stages of seeing researchers investigate the profession from quantitative and qualitative approaches.

Open source packages such as R and Talend can help investigations in qualitative or mixed-mode research using modules or components which allow text from documents or online sources to be searched for values of interest by the research team. The team seeks to determine the extent of use of the metrics-based approaches in project management and the context in which successful or failed projects may have as a root-cause an incorrect use of metrics or other considerations, such as the ability to anticipate the future, or presencing by Scharmer [3]. Using Pang and Lee [4] as a guide, along with the material from Frances and Flynn [5], the authors seek to learn how these tools help in understanding the project management profession.

References

- [1] Kerzner, H. (2009). *Project management: A systems approach to planning, scheduling and controlling*. John Wiley & Sons. Hoboken, NJ.
- [2] Lipke, W. (2009). *Earned Schedule*. Lulu publishing.
- [3] Scharmer, O. (2009). *Theory U: Learning from the future as it emerges*. Berrett-Koehler Publishers, Inc. San Francisco, CA.
- [4] Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1-2), 1–135.
- [5] Francis, L. & Flynn, M. (2010). Text Mining Handbook. In *Casualty Actuarial Society E-Forum*, 1-61.

Exploratory Analysis of 2010 Starr Hill Brewing Company Production Data Using R

Abstract:

Starr Hill Brewery, located in Crozet, VA, has increased their production over the last couple years and their products have become much more ubiquitous in the regional market. While Starr Hill continues to record measurements from every batch by hand on a datasheet, no in-process or retrospective statistical analysis is done on the data from the batches to ensure quality. The DMAIC (Define, Measure, Analyze, Improve, Control) process improvement methodology was applied to data from the year 2010 from six Starr Hill product lines to accomplish four primary goals: 1) develop suggestions for improving production efficiency; 2) reduce loss during the brewing process; 3) improve data consistency; and 4) improve product quality control. Several statistical techniques and intelligent systems algorithms were explored. This is the first year of an exploratory multi-year data mining and intelligent systems research project being conducted by an interdisciplinary undergraduate team.

Discussion of R Programming Topics

James Robison-Cox

Montana State University, Department of Mathematical Sciences
Contact info: jimrc@math.montana.edu

Keywords: Introduction to *R*, teaching *R*

This poster session is intended to engage instructors in an ongoing discussion about what (and how) we should be teaching undergraduates who have no programming skills. Several recent books target the computational complexities of such a course: [1], [3], and [2].

However, I also like to include information about graphical perception tasks from [4] and provide experience with **lattice** and **ggplot**.

I plan to create a poster which shows the similarities and differences in the topics covered by the three books cited. I will have several decks of “cards” printed with these and other topics which participants can reorder to create a preferred sequence, and I will photograph each ordering.

Hopefully discussion and idea sharing can continue after the conference via R-wiki or another website.

References

- [1] Braun, W. J. and D. J. Murdoch (2007). *A First Course in Statistical Programming with R*. Cambridge: Cambridge University Press.
- [2] Jones, O., R. Maillardet, and A. Robinson (2009). *Introduction to Scientific Programming and Simulation Using R*. Boca Raton, FL: Chapman & Hall/CRC.
- [3] Rizzo, M. L. (2008). *Statistical Computing with R*. Boca Raton, FL: Chapman & Hall/CRC.
- [4] W. S. Cleveland (1985, 1994). *The Elements of Graphing Data*. Summit, New Jersey, U.S.A.: Hobart Press.

The MDS-GUI: A Graphical User Interface for comprehensive Multidimensional Scaling applications.

Andrew Timm, Sugnet Lubbe

Department of Statistical Sciences, University of Cape Town, South Africa

Contact author A: timmand@gmail.com

Keywords: MDS, GUI, tcltk

MDS-GUI is an *R* based graphical user interface for performing Multidimensional Scaling (MDS) methods in a number of ways. The software was developed using the *R* wrapped **tcltk** package and a number of the packages affiliated to it, such as **tcltk2** and **tkrplot**. While the package is in its final stages of development, it is at this point fully demonstrable and is likely to be ready for submission to the CRAN database by the end of 2012. The intention of the **MDS-GUI** is that the menu structures and overall layout be set out in a way that is found to be user friendly and uncomplicated as well as comprehensive and effective.

The **MDS-GUI** has been developed to provide the user, even with no theoretical background on the subject, with the opportunity to perform a number of MDS methods and output a host of relevant details and graphics. In broad terms, the GUI allows the user to simply and efficiently input their desired data, choose the type of MDS they would like to perform as well as select the type of output they would like to achieve by the analysis. The use of sub-menus and property tabs gives the user the option to fine tune specific parameters of the desired MDS procedure as well as provide options to alter the way in which the resulting plots are displayed. The graphical outputs are of an interactive nature and allow the user to make adjustments to the output with a cursor to observe any difference in results. Multidimensional Scaling is usually an iterative technique, which is a quality preserved by the graphics of the software. The user is thus able to have a visual display of the processes at work and observe the moving ordination configuration.

The presentation will, first of all, provide a demonstration of the **MDS-GUI**. This demonstration will be done in such a way that highlights the features of the software from an *R* coding sense, as well as the relevance of the results from a statistical analytical point of view. In addition to this, a discussion of the development of the software will take place. This will include comments on the areas of development found to be most challenging as well as future plans for the package.

References

- [1] R Development Core Team (2011). *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [2] Grosjean, Ph. (2011). *SciViews: A GUI API for R*. *UMONS, Mons, Belgium*. URL <http://www.sciviews.org/SciViews-R>.
- [3] Luke Tierney <luke@stat.uiowa.edu> (2011). *tkrplot: TK Rplot*. *R package version 0.0-20*. URL <http://CRAN.R-project.org/package=tkrplot>

The agridat package for R

Kevin Wright¹

1. Pioneer Hi-Bred Int'l

*Contact author: kw.stat@gmail.com

Keywords: Data

The **agridat** package is a collection of about 75 data sets that have been previously published in books and journals, primarily from agricultural field experiments. The package documentation includes a description of the data and some R code that demonstrates an analysis of the data.

RMail: Automated Emailing of Sweave Reports

Dean Yergens^{1,2,*}, John Ray³, Dr. Chip Doig¹

1. University of Calgary, Canada

2. Healthcare Simulations Inc

3. Newonic Software Inc

*Contact author: dyergens@ucalgary.ca

Keywords: Reproducible Research, Sweave, Email, Automation, Reporting

We have developed a Java application to aid in the email distribution of R PDF reports written in Sweave.

This application was developed to automate the email distribution of Quality Improvement (QI) reports occurring in healthcare. We have also used the RMail application as a method for updating statistical research analysis for the Alberta Sepsis Network.

RMail is written as a stand-alone java application that installs as an executable on the Windows platform. Once launched, RMail allows a user to add a Sweave file (.Rnw) and configure settings as to when to send the out the email. These settings include Daily, Specific Weekday (Monday, Tuesday, etc), Weekdays (Monday to Friday), Monthly and Quarterly. In addition, the user can set the time of day that the Sweave report is to be executed which allows more computer intensive statistics to occur at non-peak times such as early morning (i.e. 3am). The user also has the ability to add a Title which is then used as the Email Subject. The user can also enter some default text that will be used as the Email Body. RMail also allows the user to archive all the PDFs emailed allowing for a historical record of all the Sweave PDFs generated (each PDF is appended with a timestamp). Finally, a list of email addresses can be assigned to each Sweave report.

Web-Based Epidemiology Analytic System utilizing R

Dean Yergens^{1,2,*}, John Ray³, Dr. Chip Doig¹

1. University of Calgary, Canada

2. Healthcare Simulations Inc

3. Newonic Software Inc

*Contact author: dyergens@ucalgary.ca

Keywords: Epidemiology, Reporting System, Java, RServe

One of the major strengths of *R* is the ability to interface with other programming languages allowing new applications that incorporate the ability to perform advanced statistical analysis to be developed. Several web-based applications utilizing *R* have already been reported in the literature^{1,2}.

We wish to report on a web-based analytic platform utilizing *R* and its application to the Alberta Sepsis Network (ASN) project. ASN is a provincial project investigating the epidemiology of Sepsis in Intensive Care Units (ICU). ASN incorporates a variety of data from various sources including ICU health records, administrative data (ICD10), metabolomics and other academic laboratory information.

This web-based application contains several custom developed *R* components for supporting analytics and reporting in a healthcare/epidemiology environment. The web-based analytical computing environment consists of Apache Tomcat as the server platform combined with a Postgres database backend. **RServe**³ was then used to provide an interface between our Java application running on the Tomcat server and the *R* Statistical environment.

Three distinct components utilizing *R* and **RServe** were created. The first of these components was a standard web-based reporting system that allows *R* code contained as XML to be executed and report the results back to the user's web-browser. This component also includes the ability to apply dynamic filters to the report to allow sub-setting of the data. The results in addition to being displayed in the user's web-browser can also be exported as a PDF or sent as an email.

The second component is a *R* Code repository that allows users to share their *R* code via a webpage amongst other *R* users using the system. This component includes a code window where the code can be modified and executed. Both the first and second component utilized the **RHTML** library for formatting the results produced by *R*.

The third *R* component in the web-based system is a passive dashboard system (similar to airport arrival/departure monitors) that allows for *R* charts/graphics to be displayed on a rotating basis.

This web-based system has also be utilized in several other projects including countries such as the Philippines, Zambia, Malawi and Kenya.

References

- [1] Ooms J (2010). Web development with R. *useR! 2010, The R User Conference (Gaithersburg, USA) - Abstracts of Contributed Presentations*, pp. 117.
- [2] Colombo E, et al (2010). R role in Business Intelligence Software Architecture. *useR! 2010, The R User Conference (Gaithersburg, USA) - Abstracts of Contributed Presentations*, pp. 32.
- [3] RServe - Binary R server. URL: www.rforge.net/Rserve/doc.html. Accessed: March 5, 2012

Interactive inspection of large data

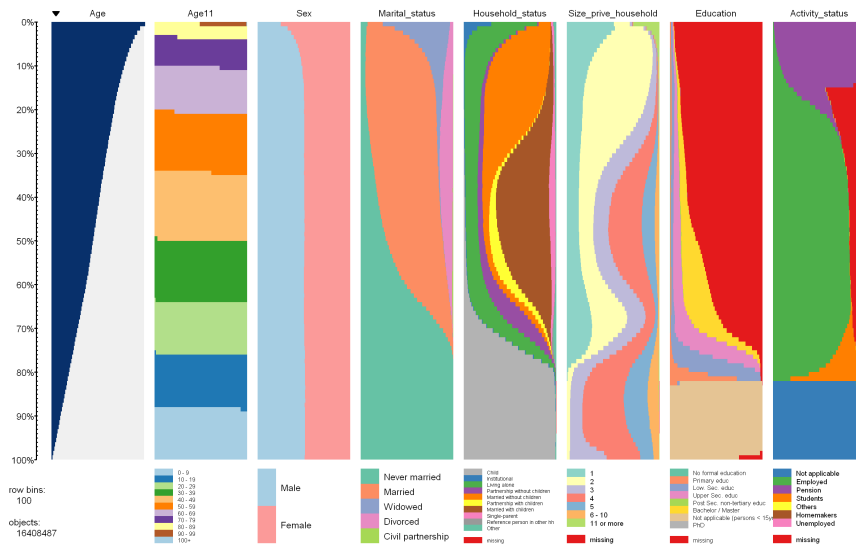
Edwin de Jonge^{1*}, Martijn Tennekes¹

1. Statistics Netherlands, The Hague/Heerlen

*Contact author: e.dejonge@cbs.nl

Keywords: multivariate analysis, visualization, large data, web application

We present **tabplotd3** an interactive tool in *R* for inspecting large multivariate data consisting of millions of observations. **tabplotd3** allows a user to visually check multivariate data for missing values, anomalies and patterns in bivariate relations, for categorical as well as numerical variables. The interface creates tableplots [1] from `data.frames` and allows to zoom in to individual observations. We applied the tool to the Dutch (virtual) census [2], creating a tableplot that depicts the 17 million inhabitants of the Netherlands.



The presentation will describe tableplots, their use and implementation in **tabplot**[3] and **tabplotd3**. The tool is a *R* package that spawns an interactive web application. To allow fast actions on multi million `data.frames` we implemented some aggregations methods in *C*. The user interface is in *javascript* and *SVG*. The tool builds on several other packages including **Rook**, **RJSONIO**, **ff** and our non interactive **tabplot**.

References

- [1] Malik, W., A. Unwin, and A. Gribov (2010). An interactive graphical system for visualizing data quality - tableplot graphics. In H. Loracek-Junge and C. Weihs (Eds.), *Classification as a Tool for Research, Proceedings of the 11th IFCS Conference*, pp. 331–339. Berlin: Springer.
- [2] Nordholt, E. S. (2005). The dutch virtual census 2001: A new approach by combining different sources. *Statistical Journal of the United Nations Economic Commission for Europe* 22(1), 25–37.
- [3] Tennekes, M., E. de Jonge, and P. Daas (2011). Visual profiling of large statistical datasets. In *NTTS, New Techniques and Technologies for Statistics*.

RSimpleITK: An R Interface to the Insight Toolkit with SimpleITK

Richard Beare^{1,2}, Daniel Blezek³, Bradley Lowekamp⁴, Brandon Whitcher^{5,6,*}

1. Monash University, Melbourne, Australia

2. Murdoch Childrens Research Institute, Royal Children's Hospital, Melbourne, Australia

3. Mayo Clinic, Rochester, MN, United States

4. National Library of Medicine, National Institutes of Health, Bethesda, MD, United States

5. Mango Solutions, Chippenham, United Kingdom

6. Imperial College London, London, United Kingdom

*Contact author: bwhitcher@mango-solutions.com

Keywords: Insight Toolkit, Medical Imaging, Registration, Segmentation

The **RSimpleITK** R package is currently in the alpha stage of development and we encourage users in the image analysis community to participate in its evolution. The functionality of both open-source environments, *ITK* and *R*, will benefit from increased communication, participation and knowledge exchange.

Highlights of the **RSimpleITK** package include:

- Automated generation of bindings via SWIG – new filters are included via rebuilding.
- Image input/output for all formats supported by *ITK* including NIfTI, ANALYZE, DICOM, TIFF, JPEG, PNG and many more.
- Representation of images and filters using *R*'s external reference mechanism, with garbage collection via *ITK*'s smart pointer infrastructure.
- Voxel access and basic image manipulation (cropping, subsampling, flipping) via *R*'s array operators.
- Straightforward import/export of image data to/from *R* arrays.
- Image arithmetic.

The Insight Toolkit (*ITK*) is an open-source software toolkit for performing registration and segmentation. Segmentation is the process of identifying and classifying data found in a digitally sampled representation. Typically the sampled representation is an image acquired from such medical instrumentation as CT (computed tomography), MRI (magnetic resonance imaging) or ultrasound scanners. Registration is the task of aligning or developing correspondences between data. For example, in the medical environment, a CT scan may be aligned with a MRI scan in order to combine the information contained in both.

ITK is implemented in C++. *ITK* is cross-platform, using the CMake build environment to manage the configuration process. In addition, an automated wrapping process generates interfaces between C++ and interpreted programming languages such as *Java* and *Python*. This enables developers to create software using a variety of programming languages. *ITK*'s C++ implementation style is referred to as generic programming (i.e., using templated code). Such C++ templating means that the code is highly efficient, and that many software problems are discovered at compile-time, rather than at run-time during program execution. It also enables *ITK* to work on two, three, four or more dimensions. A simplified interface to *ITK* that does not expose templated code, SimpleITK, is also available in multiple languages.

In order to expose a simplified version of *ITK* functionalities, SimpleITK is built internally by a code generation infrastructure comprising CMake, *Lua*, and *JSON*. A large set of commonly used *ITK* image processing filters are presented at the SimpleITK level as C++ classes that perform the equivalent image processing operations. SimpleITK provides support for 2D and 3D images, and a selected set of pixel types for them. Different image filters may support a different collection of pixel types, in many cases due to computational requirements. The library is wrapped for interpreted languages by using SWIG. In particular, the following wrappings are available: *Python*, *Java*, *Tcl*, *Lua*, *R* and *Ruby*.

Teaching Calculus with R

Daniel Kaplan^{1,*}, Randall Pruim², Nicholas Horton³

1. Macalester College

2. Calvin College

3. Smith College

*Contact author: kaplan@macalester.edu

Keywords: Calculus

Calculus is the traditional starting place for students in university-level mathematics. But rather than being an entry point to technical quantitative work, calculus is often a dead end for students. Among the shortcomings are those highlighted in the Mathematical Association of America's CUPM reports [1]: a failure to develop students' modeling and technical computation skills and a lack of connection to data and statistics. For instance, rather than addressing the weak preparation of students in technical computing, only about 20% of introductory calculus courses in the US use computers. [2]

Part of the reason for the low rate of computer utilization is an attitude that calculus is about symbolic manipulations that can be introduced by hand; part is due to cost and availability. When calculus courses do use software, common choices are *Mathematica* and *Maple*, powerful packages that nonetheless are not widely used outside of mathematics courses.

The unfortunate consequence is that a large cadre of students of relatively strong mathematical ability do not learn software skills that can be applied in other fields. And, lacking such skills, students do not learn contemporary techniques that draw on computational power.

Recognizing that statistical calculations are widely used in a broad range of fields, and that contemporary projects almost always call for the use of software for data collection, data management, and analysis, we propose to put statistics first when it comes to introducing computation in the university curriculum. To this end, we have adopted *R* for use in teaching calculus.

Our implementation, available through the **mosaic** package, provides the functionality needed for calculus: derivatives, integrals, equation solving, graphics, etc. The system uses a notation based on the `formula` class that makes it easy for students to make the translation from traditional algebraic notation and supports the transition to statistical-model building. It supports symbolic parameters, but produces conventional `function` objects that can be used in conventional ways in *R*.

Our goal is not to replace sophisticated computer algebra systems (CAS) like *Mathematica*, but to provide students with a solid basis in computing that can tie together topics that are usually taught in isolation. In addition to demonstrating the calculus capabilities of the **mosaic** package, we will describe some of the advantages of using a mainstream, non-CAS system for teaching calculus.

References

- [1] William Barker and Susan Ganter, eds. (2004) Curriculum Foundations Project: Voices of the Partner Disciplines, Mathematical Association of America
- [2] David Lutzer *et al.* (2007) Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States: Fall 2005 CBMS Survey, American Mathematical Society, p. 25

Multilevel Regression and Poststratification for Survey Data

Andrew Gelman^{1,2,3}, Michael Malecki³, Daneil Lee¹, Yu-Sung Su⁴, Jiqiang Guo³, Wei Wang^{1,*},

1. Department of Statistics, Columbia University
 2. Department of Political Science, Columbia University
 3. Applied Statistics Center, Columbia University
 4. Department of Politics, Tsinghua University
- *Contact author: ww2243@columbia.edu

Keywords: Multilevel Regression, Survey Data, Bayesian Statistics, Visualization, Spatial Statistics

Estimating public opinion on state level in US is no easy task, despite the wide spread of public opinion polls. Naive methods such as Disaggregation require enormous sample size to achieve reasonable accuracy for small states, which is either impractical in social survey data or requires pooling several inherently heterogeneous surveys. Also, it discards the demographic information, which is often very predictive of the response and available in typical survey data. Multilevel Regression and Poststratification method (mrp) provides an alternative to estimate state level (or even smaller unit) opinion. Its two-step procedure includes: (1) fit a Bayesian Hierarchical model [3] on response and demographic predictors, and (2) poststratify in reference to census data. Lax and Philips [5] showed that mrp can achieve reasonable accuracy with much less sample size (often down to the sample size of a single national survey) in estimating state level opinion on gay marriage.

We introduce an *R* package **mrp** [4] to implement the mrp method. With a target audience of social scientists in mind, **mrp** package dramatically streamlines the process of utilizing mrp method. The heavy-lifting multilevel regression part builds upon the **blme** package [2], a prior-regularized modification of `lmer` function in **lme4** package [1]. Also, by extending the `spplot` method in **sp** package [6], we enable users to intuitively and succinctly visualize the demographic variations of public opinion of all states in high quality. Besides, We take advantage of the parallelism utilities of *R* to greatly reduce the computational cost of *k*-fold cross validation for model assessment and selection.

References

- [1] Bates, Douglas, Martin Maechler, and Ben Bolker. *lme4: Linear mixed-effects models using Eigen and Eigenfaces*, 2011. <http://CRAN.R-project.org/package=lme4>.
- [2] Dorie, Vincent. *blme: Bayesian Linear Mixed-Effects models*, 2011. <http://CRAN.R-project.org/package=blme>.
- [3] Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, NY, 2007.
- [4] Gelman, Andrew, Michael Malecki, Daniel Lee, Yu-Sung Su, and Wei Wang. *mrp: Multilevel Regression and Poststratification*, 2012. <http://CRAN.R-project.org/package=mrp>. R package version 0.81-6.
- [5] Lax, Jeffrey, and Justin Philips. "How Should We Estimate Public Opinion in The States." *American Journal of Political Science* 53: (2009) 107–121.
- [6] Pebesma, Edzer J., and Roger S. Bivand. "Classes and methods for spatial data in R." *R News* 5, 2: (2005) 9–13. <http://CRAN.R-project.org/doc/Rnews/>.

Cell Lines, Chemotherapy Response, and the Need for Reproducible Research

Kevin Coombes

The University of Texas MD Anderson Cancer Center
kcoombes@mdanderson.org

Keywords: Duke University, forensic bioinformatics

In November 2006, researchers at Duke University published the first in a series of papers that claimed that (1) microarray and drug response data from cancer cell lines could be used to develop genomic signatures of response to specific chemotherapies, and (2) these signatures successfully predicted patient responses. Duke later began running clinical trials based on this work. We attempted to reproduce their analyses on publicly available data, and were unsuccessful. We identified a series of errors in data provenance and analysis, which we published as letters to editors and as an article in a statistical journal. As a result of our efforts, four clinical trials have been terminated and at least eight papers have been retracted. In this talk, I will describe some of the errors we found and some of the “forensic bioinformatics” methods used to discover them. I will also discuss tools in R and beyond that promote and enable “reproducible research.”

Flickr

Vijay Barve^{1,*}

1. Department of Geography, University of Kansas

*Contact author: vijaybarve@ku.edu

Keywords: API, Social Network Sites, Flickr

Social Networking Sites (SNS) are becoming part of daily life and has attracted attention from researchers to analyze what exactly is happening on these networks and their significance to the real world. **Twitter** package is being used widely in research projects and interesting studies ranging from Consumer attitude, mapping Twitter followers, “Twitter protests” in different parts of the world.

Flickr [1] is a popular image hosting and social networking website with more than 50 million users and 6 billion images. Users share photos and comment on photos by their friends, share photos in communities and have various types of interactions. Flickr stores date, geo-tagged location, title, description and tags (which are like keywords or labels for the photo) for each photo. About 80% photos on Flickr are public and can be viewed and the information attached to them can be accessed by anyone. As per early 2012 estimates more than 4.5m photos are uploaded per day [2]. Flickr provides Applications Programmers Interface (API) access to enable independent programmers to expand the service and access the data [3]. Flickr data is being used in various research articles on Web Technology, Natural Language Processing, Social Networks, and so on. My personal interest is to explore Flickr data store for Biodiversity related information.

Here I present *R* package to access data using Flickr APIs and some interesting applications of the data extracted from Flickr. This package uses **XML**, **rjson** and **Rcurl** packages.

References

[1] Flickr, <http://www.flickr.com/>.

[2] Yahoo Advertising Solutions - Flickr. (2012).*Yahoo! Inc.* Retrieved March 5, 2012, from <http://advertising.yahoo.com/article/flickr.html>

[3] Flickr API Documentation. (2012).*Yahoo! Inc.* Retrieved March 10, 2012, from <http://www.flickr.com/services/api/>

R, The Cloud and the Data Deluge

Karim |Chine¹

1. Cloud Era Ltd

*Contact author: karim.chine@gmail.com

Keywords: Cloud Computing, Big Data, Collaboration

Cloud computing is the answer to the explosion of big data. While the cloud provides infinite scalability for storage, several questions remain partly or fully unanswered: "How will we analyze all this data?". "How can we analyze it virtually?". "How can we leverage the programmability and elasticity of the cloud infrastructure to enhance the flexibility and capabilities of the software tools we use?". "Will we be able to produce and publish on top of models and data, analytical services and GUIs as easily as we blog?". "How will we snapshot, make reproducible, undo and redo easily data transformations and analysis?". "Will we be able to achieve software convergence and make our data analysis tools communicate and work for us in synergy?". "How will we view and analyze data collaboratively and how will we share the produced artifacts?"

Elastic-R aims to answer these questions. For the benefit of both Academia and Industry, the **Elastic-R** platform transforms *Amazon EC2* into a ubiquitous collaborative environment for data analysis and computational research. It makes the acquisition, use and sharing of all the capabilities required for statistical computing, data mining and numerical simulation easier than ever: The cloud becomes a user friendly *Google-Docs*-like platform where all the artifacts of computing can be produced by any number of geographically distributed real-time collaborators and can be stored, published and reused.

The presentation will be an overview of the **Elastic-R** platform, the latest developments will be demonstrated and applications in Bioinformatics will be illustrated.

References

- [1] Karim Chine (2010). Elastic-R Platform, <http://www.elastic-r.net>.
- [2] Karim Chine (2010). Open science in the cloud: towards a universal platform for scientific and statistical computing. In: Furht B, Escalante A (eds) Handbook of cloud computing, Springer, USA, pp 453–474. ISBN 978-1-4419-6524-0.
- [3] Karim Chine (2010). Learning math and statistics on the cloud, towards an EC2-based Google Docs-like portal for teaching / learning collaboratively with R and Scilab, icalt, pp.752-753, 2010 10th IEEE International Conference on Advanced Learning Technologies.

Scalable Embedded Scientific Computing with OpenCPU

Jeroen Ooms^{1,2,3,*}

1. UCLA Dept. of Statistics.
2. UCLA Center for Embedded Networked Sensing.
3. Stat/Dev Consulting.

*Contact author: jeroen.ooms@stat.ucla.edu

Keywords: Systems, Cloud Computing, Participatory Sensing, Web Applications, Reproducible Research

The UCLA Center for Embedded Networked Sensing (CENS) develops scalable open source systems for data collection and analysis [3]. In a current (2010-2015) project called *Mobilize*, thousands of students from high schools in Los Angeles will be armed with smartphones to collect data about their life and environment. Schools and students launch their own campaigns, and mobilize their peers to fill out surveys, make pictures and collect GPS-tagged data using an Android ‘app’ that communicates with a central server system. The students can share, explore and analyze data that they collectively gathered through web applications, or directly in *R*. The goal of the *Mobilize* project is to use participatory sensing to introduce the concept of learning from data to pre-college students in an intuitive and engaging way [1]. However, the same software is also used in other domains, like Mobile Health [2].

The *Mobilize* project illustrates how the *OpenCPU* framework [5] is used to embed *R* into scalable software systems. *OpenCPU* is a web framework on top of *rApache* [4] which defines a protocol for interacting with *R* over HTTP. It provides a layer of domain-specific functionality like object serialization, security, load balancing, resource control, reproducibility etc, while abstracting away technicalities. The *OpenCPU* protocol defines a direct mapping between *R*-functions and the RPC/REST interface. Hence, being able to write an *R* function is sufficient to publish *R* web services. This makes integration of *R* in applications and systems much easier, because the statistician only has to supply *R* code, and the web/system developer only has to call the API. *OpenCPU* has proven to enable teams of statisticians and web developers to build *R*-powered web applications, without requiring them to learn each others language.

References

- [1] Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M. (2006). Participatory sensing.
- [2] Estrin, D. and Sim, I. (2010). Open mhealth architecture: An engine for health care innovation. *Science*, 330(6005):759.
- [3] Hicks, J., Ramanathan, N., Falaki, H., Longstaff, B., Parameswaran, K., Monibi, M., Kim, D., Selsky, J., Jenkins, J., Tangmunarunkit, H., et al. (2012). ohmage: An open mobile system for activity and experience sampling.
- [4] Horner, J. (2011). *rApache: Web application development with R and Apache*.
- [5] Ooms, J. (2012). *OpenCPU: Producing and Reproducing Results*.

Tracking down bugs with bisect tools

Winston Chang^{1,2,*}

1. Rice University

2. Northwestern University

*Contact author: winston@stdout.org

Keywords: Version control, devtools, git, bug

Version control is a must for *R* program and package development. From the perspective of a version control system, a project's development history is a series of *commits*, each one making a small change to the project. If the current version of your project has a bug and you know the bug wasn't there in some previous version, then somewhere along the chain of commits between the previous and current versions, there is a place where it worked properly in one commit but didn't work in the next. That commit is the *bad* commit. If you know which commit is the bad one, it drastically reduces the search space for finding the cause of the bug; if the commit changed 10 lines of code, then the bug must be related to those 10 lines.

I will demonstrate new *bisect* tools in the **devtools** package. These tools work with the *git* version control system to do a binary search through a project's history to find the bad commit. With n commits between the previous and the current versions, the binary search can typically find the bad commit in $\log_2(n)$ steps; if there are 200 commits, you can find the bad commit in 8 steps. The bisect tools can also be used to fully automate searches for a bad commit, so you can simply write a test function and let it automatically find the bad commit.

References

Wickham, H. (2012). *devtools: Tools to make developing R code easier*, <http://CRAN.R-project.org/package=devtools>.

Dates and times made easy with lubridate

Garrett Grolemund^{1,*}, Hadley Wickham¹

1. Rice University

*Contact author: grolemund@rice.edu

Keywords: Dates, Times, Parsing, Manipulation, Time Zones

This talk presents the **lubridate** package for *R*, which facilitates working with dates and times. Date-times behave differently than other quantities, which creates technical problems for the data analyst. Chief among these are problems of *parsing*, *representation*, and *consistency*. Because date-times follow their own idiosyncratic rules, they must be identified and parsed into *R* as date-times. Parsing date-times is difficult because date-times may be represented in many ways. Formatting choices and conventions such as time zones and military times will affect how a moment of time is described and saved. **lubridate** helps analysts easily transition between different date-time representations and provides tools for easily parsing the most common formats of date-times. Modifying date-times also presents challenges. Time spans have inconsistent lengths depending on when and where they occur due to conventions such as daylight savings time, leap years, and leap seconds. **lubridate** gives an analyst the power to use or ignore these conventions with three new time span object classes for *R*. This talk will offer practical advice on how to solve date-time related problems in *R* with **lubridate**. The talk also introduces a conceptual framework for arithmetic with date-times in *R*.

TraceR: A framework for understanding *R* performance

Leo Osvald^{1,*}, Brandon Hill¹, Floréal Morandat¹, Jan Vitek¹

1. Purdue University

*Contact author: losvald@purdue.edu

Keywords: Performance, profiling, tuning, tools

Simple changes in *R* code can have dramatic performance effects. Many of these problems cannot be easily found with current *R* debugging facilities. For example, while *R*'s built-in sampling profiler can highlight long-running *R* functions, it is limited in exposing how a *R* program interacts with the interpreter. Further, it gives limited understanding of how internals, such as I/O and memory management, affect a program's performance. As users continue to push *R* to solve larger problems, understanding these performance issues becomes critical in scaling *R* programs.

We describe the TraceR suite of tools and how they can be used to examine the performance of *R* programs. This framework consists of three data collection tools built on top of *R* along with a *R* library to analyze the results. TimeR is a low-overhead counter-based profiling tool for measuring both function performance as well as the internal costs in the *R* interpreter. Once performance bottlenecks are identified, the TrackerR tool can generate detailed execution traces to better understand the exact operations occurring within trouble code. Finally, ParseR is a static analyzer for *R* code that can be used to help synchronize TrackerR results with actual source code. We will describe the use of each tool and demonstrate how they are used to understand subtle problems in real-world *R* programs.

Tools in the TraceR suite were originally designed to evaluate the overall design of the *R* language[1]. In this work, we have redesigned these tools to help *R* developers understand performance bottlenecks in their own code. These tools can now be used either within a *R* session, or applied to an entire set of programs. This allows for either interactive single program analyses or analyzing how a library performs across a set of programs.

References

- [1] Morandat, F., B. Hill, L. Osvald, and J. Vitek (2012). Evaluating the Design of the R Language. In *ECOOP 2012, 26th European Conference on Object-Oriented Programming (Beijing, China)*.

Helping Your Organization Migrate to R

Robert A. Muenchen ^{1*}

1. University of Tennessee

*Contact author: muenchen@utk.edu

Keywords: r-project, *SAS*, *SPSS*, system migration

The use of *R* is growing rapidly, taking an ever increasing proportion of the analytics market [1]. Transitioning an organization from proprietary systems such as *SAS* or *SPSS Statistics* can provide greater functionality while saving thousands of dollars in annual license fees. However system migrations also require significant resources. This talk will describe steps to consider when planning such a migration including: motivation, training, documentation, conversion services, phased migration, choosing packages that provide familiar output and considering the impact on both style of work and where *R* fits into people's complete research work flow.

References

[1] Muenchen (2012). The Popularity of Data Analysis Software, <http://r4stats.com/articles/popularity>

Graphical tool for disambiguation of bias in health research

Drew Griffin Levy^{1*}, David Norris²

1. Genentech, Inc.; 2. David Norris Consulting, LLC

*Contact author: levy.drew@gene.com

Keywords: Bias, Epidemiology, R programming, Statistical analysis, Statistical graphics

Health research is frequently concerned with evaluating the relationship between a risk factor or therapeutic exposure of primary interest (Tx) and a health outcome. For non-randomized (“observational”) studies the analyst must be concerned with non-causal association between the Tx and outcome inadvertently due to the effects of other variables: a.k.a “confounding”. Bias in the estimate of effect of Tx on outcome due to confounding is a function of both (i) the association between ancillary variables and the outcome, and (ii) differential distribution of the ancillary variables across levels of Tx. The magnitude of bias due to confounding in an estimate of the Tx effect is an aggregate of the confounding effects of all ancillary variables.

Well designed graphs cogently reveal structure and patterns for specific analytic purposes. Radar plots are a useful way to display multivariate data with an arbitrary number of variables. We elaborate on the graphical idiom of the radar plot to express the multivariate relations among variables in an analysis of health outcomes data for the purpose of elucidating the potential in aggregate for bias. In this application of the radar plot the radial axes denote covariates in the analysis (e.g., baseline patient attributes, prognostic indicators, concomitant medications, etc.). Each radii is scaled and oriented with the range minimum at the origin and the range maximum at the outer terminus. Intersecting the axes are colored lines denoting comparison (Tx or exposure) groups. Where these transecting lines fall on radii indicates the point estimate for the statistical summary for each group, making comparison of values for multiple variables within a Tx group easy, and facilitating the perception of multivariate differences between the groups of interest. The quantitative information for the covariate axes are summarized in the following manner: continuous variable as a median, dichotomous variable as a proportion, an ordinal variable as the mean of numeric integers.

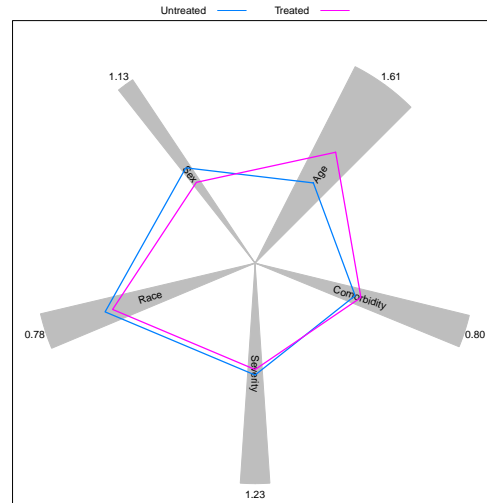


Figure 1: Bias-potential graph

Overlying each covariate axis is a radial polygon whose size is proportional to the magnitude of the odds ratio between the covariate and outcome. This feature assists in comparing the relative importance of covariates for potential bias. Multi-panel conditioning is incorporated into the function for stratifying the data and comparing subgroups.

It is proposed that this *bias-potential graph* does a more comprehensive job of representing the complex of associations that contribute to potential confounding in an observational study by simultaneously addressing all components of the conditions for confounding. Other applications of this graphic may include the elucidation of selection bias, treatment propensity, randomization balance, and evaluation of success of propensity score adjustment.

This *bias-potential graph* is generated by a flexible function written in R, and can take advantage of features of the **Hmisc** and **rms** packages. The *bias-potential graph* can thereby be integrated into analysis and reporting within the R environment.

Decision Making under Uncertainty: R implementation for Energy Efficient Buildings

Emilio L. Cano^{1,2,*}, Javier M. Moguerza¹

1. University Rey Juan Carlos

2. University of Castilla-La Mancha

*Contact author: emilio.lopez@urjc.es

Keywords: Decision Making, Optimization, Stochastic Programming, Energy Efficiency, Modeling

Decision making is often made under uncertain conditions. Optimization problems are usually formulated as mathematical programming models, where a series of parameters are fixed. Usually some of these parameters are unknown and they are estimated, using this estimation as a known parameter. This way of solving the problem returns the best solution for the expected value of each parameter. Most times, this solution is not optimal for the actual value of the parameters, and even it can be really inaccurate. A more effective way to cope with uncertainty in optimization problems is Stochastic Programming. Stochastic Programming finds the optimal solution for the optimization problem, taking into account the probability distribution of the stochastic parameters, that is to say, not only their expected values, but also their variability. In fact, the solution obtained is not an optimal solution for a specific value of the stochastic parameters, but it is a common solution that takes into account all the possible scenarios. Stochastic Programming has been used for decision making in energy markets (e.g. [2]). Within the EnRiMa project¹ a Decision Support System is being developed to support building managers in optimizing both strategic and operational decisions. There have been many optimisation approaches to this problem, of which *DER-CAM*² [4, 5] is one. Complex systems as the ones in the EnRiMa project need to be accurately described in a condensed way representing the huge amount of variables, parameters and constraints in the models. A *Symbolic Model Specification* (SMS) has been developed using *R*, providing an integrated framework for symbolic models and implementation. Using customized functions for the manipulation of data structures, we can easily build the SMS equations in \LaTeX documents using *Sweave*. The models can be solved with *R* (e.g. using *lpSolve* [1]), or we can generate the adequate files for specialized software such as *GAMS* [3]. We acknowledge projects EnRiMa (FP7 project 260041) and RIESGOS-CM (code S2009/ESP-1685), in which the methodology described in this work has been applied.

References

- [1] Berkelaar, M. and others (2011). *lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs*. R package version 5.6.6.
- [2] Conejo, A., M. Carrión, and J. Morales (2010). *Decision Making Under Uncertainty in Electricity Markets*. International Series in Operations Research and Management Science Series. Springer.
- [3] GAMS (2012). *gdxxrw: interfacing gams and R*. Internet. retrieved 2012-03-06.
- [4] Marnay, C., J. Chard, K. Hamachi, T. Lipman, M. Moezzi, B. Ouaglal, and A. Siddiqui (2001). Modeling of customer adoption of distributed energy resources. Technical report, Lawrence Berkeley National Laboratory.
- [5] Siddiqui, A. S., C. Marnay, J. L. Edwards, R. Firestone, S. Ghosh, and M. Stadler (2005). Effects of carbon tax on microgrid combined heat and power adoption. *Journal of Energy Engineering* 131(1), 2–25.

¹Energy Efficiency and Risk Management in Public Buildings. <http://www.enrima-project.eu>

²Distributed Energy Resources – Customer Adoption Model

Risk Management: An Econometric Analysis of Risk in Energy Market

Emmanuel Senyo Fianu^{1,2,*}, Luigi Grossi^{1,1}

1. Department of Economics, University of Verona, Italy and Department of Economics, University of Verona, Italy

2. Department of Business and Economics, Aarhus University

*Contact author: emmanuelsenyo.fianu@univr.it

Keywords: Electricity prices, Model Confidence Set, Risk management, AR–GARCH, Extreme Value theory.

The energy market specifically electricity market all over the world is going through a great transition. From being a regulated market with no or very low uncertainty in future earnings, the market is now becoming liberalised and deregulated. The prices of electricity are no longer determined by regulator but by market participants. Price fluctuation and partial comovement with demand are a feature inherent in the liberalised electricity market. The most vital test for the new market regime is its ability to manage the excessive volatility inherent in a system with substantial capacity variations from year to year and from season to season. This newly created climate requires protection against market risk and has become very essential. In this paper, we proposed an AR–GARCH–type–EVT with various innovations and their skewed variants based Value at Risk and Conditional Value at Risk for electricity price risk quantification for different emerging electricity market. Value at Risk gives an estimate for the maximum daily electricity price change associated with a confidence (likelihood) level, with conditional Value at Risk as an alternative risk measure and provide good source of information in designing risk management strategies. We therefore carry out risk analysis on energy market using **rugarch** package in R programming language. Our findings suggest that there is no correct exceedances for the out of sample Value at risk performance and hence Extreme Value Theory approach has been adopted parametrically to compute the value at risk and the conditional value at risk for the various market under study. This paper applies the Model Confidence Set (MCS) procedure of Hansen, Lunde, and Nason (2003) to the different models which selects the best model with a given level of confidence.

References

- [1] Ghalanos, A. (2011). *rugarch: Univariate GARCH models*. R package version 1.0-5.
- [2] Hansen, P. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of Applied Econometrics* 20(7), 873–889.

Social Network Analysis for Online Group Games

Jackie McCush Jr.¹

1. Analytics and Insight, Quaero as CSG Solution, CSG International
*Contact author: jack.mccush@csgi.com

Keywords: social network analysis, business analytics, online, data mining

Companies that build online group games have learned that members of these groups act as a Social Network. Most of these companies have customers that play multiple games with different members. There is tremendous value in understanding the group dynamics that can affect how many people will play in a group, how long they will play, how often they play. In this work I will show how to use *R* to analyze these networks and calculate measures of density, components, cliques, centralization, factions, centrality closeness and betweenness. Once the network is drawn I will discuss how identifying influencer can be turned into marketing programs such as loyalty or retention.

References

- [1] Conway, Drew (2009). Social Network Analysis in R, http://www.drewconway.com/zia/wp-content/uploads/2009/08/sna_in_R.pdf.
- [2] McMahan, Peter; Moody, James; White, Doug (2003). Cohesive Blocking. http://intersci.ss.uci.edu/wiki/index.php/Cohesive_blocking
- [3] Chapsky, Daniel (2011). Leveraging Online Social Network Data and External Data Sources to Predict Personality. In useR! 2011, The R User Conference (Warwick, United Kingdom), pp. 31–37.

General Process Data Management Framework

R.F. Rossouw^{1*}, R.L.J. Coetzer¹, A.G. Mostert¹

1. Sasol Technology Research and Development, PO Box 1, Sasolburg, 1947, South Africa.

*Contact author: ruan.rossouw@sasol.com

Keywords: Online Data, Web Interface, Key Performance Indicators

In the Petro-Chemical Industry large volumes of data are generated daily. These data are valuable resources and it is important to make efficient use of the data. The raw data is stored in different formats and technologies. For the end user to make efficient use of the data it must be easily accessible and in a convenient format. Specifically, the end user should not be aware of the underlying computer codes and technologies. In this paper a General Framework for Process Data Management will be presented. This framework entails a combination of *R*, *C*, *ODBC*, *VBA*, *MySQL*, *PHP*, *JavaScript* and *HTML*. The end user interacts with the framework via a web interface, or Excel Add-In. The developed framework has been implemented in Sasol for on-line key performance indicator monitoring.

Using R to Evaluate Quality Auditor Consistency for a Call Center

Samuel Thomas¹

*Contact author: samuel.thomas522@gmail.com

Keywords: categorical, visualization, chi-square, business intelligence

Auditing calls is a critical element of quality monitoring. Many call centers have established quality monitoring guidelines, which independent auditors are supposed to follow when evaluating calls.

Many hours of planning and calibration can go into designing a quality monitoring system and then training an auditing team to implement the system. In an ideal world, every auditor would interpret the scoring guidelines the same way, thereby scoring each call exactly the same way.

Of course, we do not live in an ideal world. Even the most specific of guidelines can be open to interpretation when exposed to real-world calls.

Still, we need auditors to follow the guidelines to a certain degree of consistency or quality scoring loses its credibility. How can we assess whether scoring differences by auditors are due to randomness, bias, or other factors?

Fortunately, statistical tools are available to help make such determinations. One such tool is the chi-square test of independence. This test can be used to determine whether there is evidence of auditor bias or not.

This test and corresponding data visualization are freely available from an open source statistical software program called R.

The plots and statistical tests used are based on functions from a contributed R package called **vcd**, which stands for “Visualizing Categorical Data.”[1]

References

- [1] Hornik, K., A. Zeileis, and D. Meyer (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software* 17.

Spatio-temporal analysis of climatic data in understanding species' distributions.

Narayani Barve^{1*}

1. Biodiversity Institute, Ornithology division, University of Kansas, Lawrence, KS - 66045

*Contact author: narayani@ku.edu

Keywords: Ecological niche / Species Distribution modeling, ERA-reanalysis climatic data, Spanish Moss, ncdf, raster

Understanding species' distributions is a major challenge in biodiversity science, conservation biology and biogeography (2,3). The emerging field of ecological niche modeling uses species' occurrences and climatic data to estimate potential distributions of species across landscapes. However, potential distributions generated from these correlational models do not consider physiological tolerances explicitly, which represents a potentially very serious limitation (1). To provide a first view of the magnitude of this problem, I integrated physiological thresholds of Spanish Moss measured at one location over the known geographic distribution of Spanish Moss. Spanish Moss is distributed from South America to southeastern United States (4). Climate parameters used include minimum and maximum temperatures, and rainless days (5,6). To generate the physiologically-based maps, I processed 6 hourly daily environmental data of 1.5 degrees developed by ERA interim reanalysis, over 1989-2010 to generate maps of potential distribution of Spanish Moss variable by variable. Daily environmental data is downloaded from ERA interim website, which is stored in NetCDF format. Using **ncdf** and **raster** package in R programming language, a R script is written to process the data, which in turn gave a suitability map as per the threshold for a variable. In theory, the intersection of these suitable variable maps should yield the area where Spanish Moss can maintain populations; I present detailed comparisons of these areas with known occurrences of the species hemisphere wide, and analyze the factors that produce agreement and disagreement among the two views.

References

- [1] Araújo, M. B., and Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), 1677-1688.
- [2]. Darwin, C. (1859). On the Origin of Species. *J. Murray, London*.
- [3]. Gaston, K. (2003). The Structure and Dynamics of Geographic Ranges. *Oxford University Press, Oxford*.
- [4]. Garth R.E. (1964). The Ecology of Spanish Moss (*Tillandsia Usneoides*): Its Growth and Distribution. *Ecology*, 45(3), 470-481.
- [5]. Martin, C. E., Christensen, N. L., and Strain, B. R. S. (1981). Seasonal Patterns of Growth, Tissue Acid Fluctuations, and ¹⁴CO₂ Uptake in the Crassulacean Acid Metabolism Epiphyte *Tillandsia usneoides* L. (Spanish Moss). *Oecologia*, 49(3), 322-328.
- [6]. Martin, C. E., and Siedow, J. N. (1981). Crassulacean Acid Metabolism in the Epiphyte *Tillandsia usneoides* L. (Spanish Moss): Responses of CO₂ exchange to controlled environmental conditions. *Plant physiology*, 68(2), 335-9.

Modeling North American vegetation: solving a 46-level classification problem

Nicholas L. Crookston*¹, Gerald E. Rehfeldt¹

1. USDA Forest Service, Rocky Mountain Research Station

*Contact author: ncrookston.fs@gmail.com

Keywords: Random Forest

We describe the (perhaps novel) approach we used to solving a 46-level classification problem using *R* package **randomForest**, which has a limit of 32 levels. Our approach used an ensemble of 100 forests that each modeled 9 of the 46 levels plus a 10th class that was coded “other”. The 9 levels selected for each of the 100 forests were selected using combination of three strategies designed to make the ensemble an effective predictor. Climate variables were used as the predictors [1]. An ensemble predictor was built that used all 100 forests. This predictor was free to predict the class “other” as the most likely class level signaling that, given the values of the predictor variables for a case at hand, the most likely class was “none of the others.” This outcome prevailed when novel future climates were presented to the ensemble classifier as new data which was nearly never the case for contemporary climate but frequently the case for (a) contemporary climates from outside North America and (b) climates predicted for the end of the century inside North America. Predictions for each ~1km pixel (0.00083 arc seconds) of North America were produced for contemporary and future climates. Maps depicting spatial distribution of biomes now and in the future were produced. Notably, overlays that depict uncertainty due to the presence of novel conditions were possible using this ensemble approach [2]. We conclude that the approach would likely be useful in other disciplines.

References

- [1] Climate data are available at, <http://forest.moscowfsl.wsu.edu/climate/>
- [2] Rehfeldt, G.E.; Crookston, N.L.; Senz-Romero, C.; Campbell, E.M. (2012). North American vegetation model for land-use planning in a changing climate: a solution to large classification problems. *Ecological Applications* 22(1):119-141.

Rquake: Earthquake Hypocenter Analysis

Jonathan M. Lees^{1,*}

1. University of North Carolina, Chapel Hill

*Contact author: jonathan.lees@unc.edu

Keywords: Non-linear Inversion, Graphics, Geophysics, Earthquakes

Rquake, is new package dealing with all aspects of seismology and earthquake behavior, recently installed on CRAN. **Rquake** includes code for non-linear, earthquake hypocenter determination from arrival times estimated on local seismic networks. The package includes a graphical interface showing spatial convergence of hypocenter locations as a function of SVD and Tikhonov regularization. Iterative and robust inversion procedures are incorporated to reduce noise and localized station effects. Automated arrival time triggering and event association can be applied to provide initial focal parameters, although in volcano and geothermal settings these are often difficult to implement. Hypocenter error ellipsoids are produced illustrating uncertainty in locations and covariance of spatial parameters. Inversion for one-dimensional velocity models is currently being developed, although the **Rquake** package will ultimately include a full, three-dimensional tomographic inversion. The graphical interface utilizes and complements the **RSEIS** and **GEOmap** packages, as well as the **RFOC** package for focal mechanisms, providing a comprehensive seismology platform for analysis of local earthquake and seismicity patterns. Examples from geothermal networks in California, Ecuador and Chile will illustrate the interactive procedures that allow users to optimize parameters associated with seismic analysis.

rgeos: spatial geometry predicates and topology operations in R

Colin Rundel^{1,*}, Roger Bivand², Edzer Pebesma³

1. Duke University, Department of Statistical Science
 2. Norwegian School of Economics, Department of Economics
 3. University of Münster, Institute for Geoinformatics
- *Contact author: rundel@gmail.com

Keywords: Geospatial, GIS, geometry, sp

rgeos is a package that implements functionality for the manipulation and querying of spatial geometries using the Geometry Engine - Open Source (GEOS) C library. This package expands on existing spatial functionality in R through integration with **sp** spatial classes and transparently replaces **gpclib** which is encumbered by a licensing agreement allowing only non-commercial use. Additionally, **rgeos** includes functionality for spatial predicates and topology operations for non-polygon geometries like points, lines, linear rings, and heterogeneous geometry collections. Previously, these operations were not possible within R and required the use of external GIS tools like GRASS, PostGIS, or ArcGIS which significantly complicate spatial workflows.

This talk will discuss the basic usage of this package with a focus on real world use-cases from the R-sig-Geo mailing list. Additionally, we will cover some of the finer details of the GEOS library which will give insight into the most efficient approaches for employing **rgeos**. Finally, we will discuss future directions for the package with plans for additional features such as spatial indexes using GEOS' `STRtree` functionality and the addition of improvements made in GEOS 3.3.0.

Generation of synthetic universes for micro-simulations in survey statistics

Jan-Philipp Kolb¹

1. GESIS - Leibniz Institute for the Social Sciences
Survey Design and Methodology
P.O. Box 122155
68072 Mannheim

Keywords: Synthetic data generation, comparative simulation studies, survey statistics

Observational data is a necessary basis of research in social sciences and humanities. But for many researchers it is only possible to access this data via controlled remote data processing or on-site usage. Especially for *R*-users it is helpful to try out the programs before the usage of these expensive possibilities.

Micro-simulation is often used to control for the interplay between data structure, sampling scheme, and properties of estimators. But often the available data is insufficient to run a simulation study. This may be attributed to disclosure reasons.

Rubin [7] published an idea to encompass the problem of disclosure risk. It is the generation of synthetic datasets from existing confidential survey data.

The aim of this work is it to provide methods for the generation of a synthetic universe based on the work done for the AMELI project which was funded under the 7th FWP (Seventh Framework Programme) of the European Commission. This may be used to test methods in a simulation study within a design-based environment.

The requirements for such a population are manifold, therefore methods are presented to test for disclosure risk and the reproduction of the most important population characteristics.

References

- [1] Alfons, A., P. Burgard, P. Filzmoser, B. Hulliger, J.-P. Kolb, S. Kraft, R. Münnich, T. Schoch, and M. Templ (2011). The ameli simulation study. Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI.
- [2] Alfons, A., P. Filzmoser, B. Hulliger, J.-P. Kolb, S. Kraft, R. Münnich, and M. Templ (2011). Synthetic data generation of silc data. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
- [3] Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to eu-silc. *Statistical Methods & Applications* 20, 1–25.
- [4] Drechsler, J., A. Dundler, S. Bender, S. Rässler, and T. Zwick (2008, December). A new approach for disclosure control in the iab establishment panel - multiple imputation for a better data access. *ASIA Advances in Statistical Analysis* 92(4), 439–458.
- [5] Münnich, R. and J. Schürle (2003). On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen.
- [6] Raghunathan, T., J. Reiter, and D. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19, 19:1–16.
- [7] Rubin, D. (1993). Discussion on statistical disclosure limitation. *Journal of Official Statistics* vol. 9, pp. 461–468.
- [8] Voas, D. and P. Williamson (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography* 6, 349–366.

Interactively Mapping Significant Differences Between Areas

Jerzy Wieczorek^{1*}

1. U.S. Census Bureau

*Contact author: jerzy.wieczorek@census.gov

Keywords: Interactive graphics, Mapping, Uncertainty, Visualization

Geographers have proposed several ways to convey uncertainty and compare the statistical quality of the area-level estimates they are mapping in choropleth maps [1]. One idea is to ensure that significantly-different estimates have different colors or shadings. In this approach, the colors ought to indicate the significances of all pairwise comparisons at once. This is not always practical or even possible, much less easy [2]. We present an alternative approach that uses the **mmaps** package in *R* to display significance of differences between areas in choropleth maps, color-coded relative to one (interactively-selected) baseline area at a time. This is similar to another Census Bureau effort [3], for mapping American Community Survey data in ArcGIS. However, whereas that effort is a plugin to a proprietary tool and focuses on a specific dataset, the R code presented in this talk is flexible enough to use with any dataset that includes estimated variances or margins of error. We illustrate the approach using state and county data from the Small Area Income and Poverty Estimates program.

References

- [1] Harrower, M. (2003). Representing Uncertainty: Does it Help People Make Better Decisions? <http://www.ucgis.org/Visualization/whitepapers/Harrower.pdf>.
- [2] Francis, J., J. Vink, N. Tontisirin, S. Anantsuksomsri, and V. Zhong (2012). Alternative Strategies for Mapping ACS Estimates and Error of Estimation, http://www.cpp.cornell.edu/html/events/CPC_Strategies%20for%20Mapping%20ACS%20Estimate%20and%20MOE.pdf.
- [3] Torrieri, N., D. Wong, and M. Ratcliffe (2011). Mapping American Community Survey Data. In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association, pp. 1089–1100.

New data correction features of `editrules` and `deducorrect`

Mark van der Loo^{1*} and Edwin de Jonge¹

1. Statistics Netherlands, PO-box 2490 HA The Hague, the Netherlands
*Contact author: m.vanderloo@cbs.nl

Keywords: Automated data cleaning, numerical, categorical, and mixed data

Before raw data can be transformed to meaningful statistical output, statisticians often have to spend considerable effort preparing a dataset for analyses. Such preparations include repairing inconsistencies and imputing missing data. Packages `editrules` and `deducorrect` are developed to help statisticians to increase data quality by offering convenient ways to define and manipulate the rules that a dataset must obey. These rules include record-wise linear (in)equations and logical restrictions. Moreover, the rules can be put to use to check datasets, localize erroneous fields, repair common errors and derive missing values with certainty, where possible.

Over the last year, the packages have received around 300 `git` commits. In our presentation we will highlight some of the new features, which include

- Support for logical constraints on categorical and mixed data, *e.g.*

```
if ( age == "under-aged" ) !married
if ( nEmployees > 0 ) salaryPayed > 0.
```
- Reading numerical, logical and mixed constraints from a free-format file.
- Increased speed for error localization through mixed integer programming.
- Impute missing values when they can be uniquely derived from observed values using numerical or logical constraints.
- Visualisations of restriction graphs, restriction violations, error locations.

We will also touch upon the internal representation of restrictions and advise on a general work flow when using these packages. Future developments will be discussed as well.

Examples, theory and algorithms behind the packages can be found in the references. All references are included as documentation with the packages, which are available via the Comprehensive R Archive Network.

References

- [1] De Jonge, E. and M. van der Loo (2011a). Error localization as a mixed integer problem with the `editrules` package. Technical report, Statistics Netherlands, The Hague. Forthcoming.
- [2] De Jonge, E. and M. van der Loo (2011b). Manipulation of linear edits and error localization with the `editrules` package. Technical Report 201120, Statistics Netherlands, The Hague.
- [3] Van der Loo, M. and E. de Jonge (2011a). Deductive imputation with the `deducorrect` package. Technical Report 201126, Statistics Netherlands, The Hague.
- [4] Van der Loo, M. and E. de Jonge (2011b). Manipulation of categorical data edits and error localization with the `editrules` package. Technical Report 201129, Statistics Netherlands, The Hague.
- [5] Van der Loo, M., E. de Jonge, and S. Scholtus (2011). Correction of rounding, typing and sign errors with the `deducorrect` package. Technical Report 201119, Statistics Netherlands, The Hague.

Using R with node.js as a web service

Joe Olson^{1*}, Cory Nissen^{1*}

1. Akoya Inc.

*Contact author: useR2012@akoyainc.com

Keywords: service, node, socketConnection, socket

A popular use for *node* is to consume streaming data. Streaming data sources include health data, radio wave data, twitter feeds, location tracking data, etc. *Node* provides an easy way to consume such data, and also to serve data, but it does not offer robust real time analytics. That's where *R* comes in. Communicating with *node* on the same machine or a separate machine via a socket connection, *R* can provide the missing real time analytics piece without the complications of setting up and using **Rserve** or **rApache**. Both **rApache** and **Rserve** are great packages when the intent is to use *R* as an “all-in-one” web service package, but our preference was to use *node* for it's strengths in stream consumption and basic web serving and *R* for it's analytical strengths.

Web Applications for Statistics in Quality Management

Thomas Roth*, Johannes Schober

The Department of Quality Science - Technical University of Berlin

*Contact author: thomas.roth@tu-berlin.de

Keywords: rApache, Sweave, Measurement Systems Analysis, Statistical Process Control, Design of Experiments

Some of the reasons for the success of *R* are its availability and community adding so much value to *R* using packages. Despite this convenient availability of *R* for most platforms, scenarios can be identified where the use of (specific) statistical methods, requiring a web browser only and thus maximizing the aspect of availability, would be pleasant. Such scenarios can be found in corporations along with the need for reporting functionalities but also in teaching statistics, e. g. to engineers, where an installation of *R* might not be feasible.

An approach is presented, providing specific statistical procedures [4, 7] of *R* and commonly required report generation used in Quality Management and attributed to the Six Sigma methodology [2], in the form of a web application using rApache [1], Sweave [3] and complementary techniques. This approach is illustrated by examples regarding typical topics such as Design of Experiments, Measurement Systems Analysis or Statistical Process Control.

References

- [1] Horner, J. (2011). *rApache: Web application development with R and Apache*.
- [2] ISO (2011). Quantitative methods in process improvement – Six Sigma – Part 2: Tools and techniques (ISO 13053-2:2011).
- [3] Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), *Compstat 2002 — Proceedings in Computational Statistics*, pp. 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.
- [4] Roth, T. (2011). *qualityTools: Statistics in Quality Science*. R package version 1.50 <http://www.r-qualitytools.org>.
- [5] Roth, T. and J. Herrmann (2010). Teaching Statistics in Quality Science using the R-Package qualityTools. In *useR! 2010, The R User Conference (Gaithersburg, Maryland USA)*, pp. 137.
- [6] Roth, T. and R. Jochem (2012). Web Applikationen im Qualitätsmanagement (mit R). In R. Woll (Ed.), *Vielfalt Qualität - Tendenzen im Qualitätsmanagement*, Volume 13 of *Berichte zum Qualitätsmanagement*, pp. 61–75. Shaker.
- [7] Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. *R News* 4/1, 11–17.

Creating interactive web pages within *R*: The **gWidgetsWWW2** package

John Verzani^{1*}

1. CUNY/College of Staten Island *Contact author: verzani@math.csi.cuny.edu

Keywords: Web programming, GUIs, Interfaces

The **gWidgetsWWW2** package allows *R* programmers to easily create interactive web pages within the *R* programming language, without needing a knowledge of *HTML*, *JavaScript* or a server-side scripting language. The package implements the **gWidgets** API for writing graphical user interfaces within a *JavaScript* library (Ext JS). This is an easy to learn API that, though stripped down, provides enough flexibility for many applications. These web pages can be deployed locally through the *R*'s help server with the **Rook** package or can be served remotely with the **FastRWeb** package. While not a solution for scaled-up usage, the ease of rapid prototyping and avoidance of web-technology know how for the programmer makes it a useful tool for the right job. The presentation will cover the basics of designing and implementing an interactive interface within the framework

Parallel R, Revisited

Norman Matloff

University of California, Davis
*Contact author: matloff@cs.ucdavis.edu

Keywords: parallel computation, multicore, GPU, embarrassingly parallel

As algorithms become ever more complex and data sets ever larger, *R* users will have more and more need for parallel computation. There has been much activity in this direction in recent years, notably the incorporation of **snow** and **multicore** into a package, **parallel**, included in *R*'s base distribution, and the various packages listed on the CRAN Task View on High-Performance and Parallel Computing. Excellent surveys of the subject are available, such as (Schmidberger, 2009). (Note: By the term *parallel R*, I am including interfaces from *R* to non-*R* parallel software.)

Yet two large problems loom: First, parallel *R* is still limited largely to “embarrassingly parallel” applications, i.e. those that are easy to parallelize and have little or no communication between processes. Many applications, arguably a rapidly increasing number, do not fit this paradigm.

Second, the platform issue, both in hardware and software support aspects, is a moving target. Though there is much interest in GPU programming, it is not clear, even to the vendors, what GPU architecture will prevail in the coming years. For instance, will CPUs and GPUs merge? Major changes in the dominant architecture will likely have significant impacts on the types of applications amenable to computation on GPUs. Meanwhile, even the support software is in flux, especially omnibus attempts to simultaneously cover both CPU and GPU programming. The main tool currently available of this sort, OpenCL, has been slow to win users, and the recent news that OpenACC, a generic GPU language, is to be incorporated into OpenMP, a highly popular multicore language, may have negative implications for OpenCL. These fast-moving and unpredictably-directed trends in hardware and software make it difficult for *R* to parallelize.

In this talk, I will first discuss the above issues, and then offer two approaches to partly deal with them. The first approach is algorithmic, applicable to general statistical estimators (functions of i.i.d. sample data). Here I replace what typically will be non-embarrassingly parallel computations by embarrassingly parallel asymptotic equivalents. (The asymptotic nature will be seen not to be an issue, since problems that are large enough to need parallel computing will be well past needed convergence levels.)

Second, I will introduce a new *R* package, **Rth**, which will be an *R* interface to Thrust. The latter is in turn an interface NVIDIA provides for CUDA GPU coding, but for which the user also can also take multicore CPU as the backend. Unlike CUDA, OpenCL, OpenMP and the like, Thrust operates at a high level, offering operations as sorting, reduction, prefix sum, search and so on, all usable either on GPUs or multicore CPUs (producing either CUDA or OpenMP code). **Rth** will thus provide *R* users will easy access to these operations in a cross-platform manner, much like Magma does for matrices. I will argue that this kind of hybrid approach (if not this particular implementation) may enabling the *R* community to “hedge their bets” in the face of the uncertain hardware situation.

References

M. Schmidberger *et al* (2009). State of the Art in Parallel Computing with *R*. *Journal of Statistical Software*, 1–27.

New Tools for Reproducible Research with R

Yihui Xie^{1,*}, J.J. Allaire²

1. Department of Statistics, Iowa State University

2. RStudio

*Contact author: xie@yihui.name

Keywords: Reproducible, Sweave, knitr, RStudio, LyX

The practice of reproducible research (Schwab et al. [4]) seeks to make scholarship and data analysis more transparent, verifiable, recreateable, and reliable. *R* includes a rich set of tools for reproducible research based on Sweave (Leisch [1]), a system for dynamic document authoring.

In this talk we will discuss and demonstrate several new tools that have the potential to make reproducible research with *R* more flexible, powerful, and accessible to a broader audience. The first of these is the **knitr** package (Xie [6]), which follows in the steps of Sweave and adds several new capabilities including caching, tikz graphics, and code highlighting and reformatting. There are also features motivated by feedback from Sweave users such as conditional evaluation and code externalization. Graphics capabilities have also been significantly enhanced. This includes caching, displaying multiple plots per chunk, control over individual plot sizes, flexible arrangement (side by side or even as animations), and compatibility with non-standard plots from other packages like **rgl** and **rggobi**. Knitr supports traditional Sweave documents as well as provides several built-in output hooks which can be used to write results in \LaTeX , HTML and Markdown.

One of the difficulties in persuading users of traditional authoring environments to move to a reproducible workflow is the availability of tools that are strong for both document editing and computation. We'll demonstrate some new capabilities of the RStudio integrated development environment (RStudio [3]) designed to bridge this gap. RStudio combines TeX authoring features such as spell-checking, error navigation, and smart PDF previewing with Sweave and knitr savvy code editing tools. We will also demonstrate the support for Sweave and knitr that have been added to the LyX document editor (LyX [2]).

One of the cornerstones of reproducible research is authoring in plain-text formats. Recently several lightweight markup languages (Wikipedia [5]) such as Org-mode, Markdown, and AsciiDoc have emerged and have the potential to make authoring reproducible research more straightforward and accessible to a much broader audience. We'll discuss these new formats and the various ways that they can be used with *R*.

References

- [1] Leisch, F. (2002). *Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis*. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.
- [2] LyX (2012). LyX 1.6.1 - The Document Processor. <http://www.lyx.org>.
- [3] RStudio (2012). Rstudio: Integrated development environment for r. <http://www.rstudio.org>.
- [4] Schwab, M., M. Karrenbach, and J. Claerbout (2000). Making scientific computations reproducible. *Computing in Science Engineering* 2(6), 61–67.
- [5] Wikipedia (2012). Lightweight markup language. http://en.wikipedia.org/wiki/Lightweight_markup_language.
- [6] Xie, Y. (2012). knitr: A general-purpose package for dynamic report generation in r. <http://yihui.github.com/knitr/>.

Parallel data processing with R on combined IBM Netezza 1000 and HPC cluster

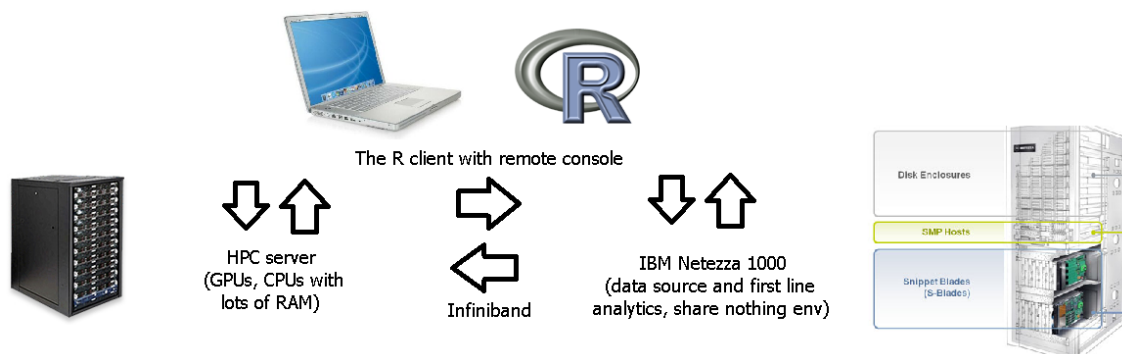
Przemyslaw Biecek*, Grzegorz Maj, Ala Strachocka,
Pawel Chudzian, Maciej Michalewicz, David Flaxman

IBM Netezza R&D Labs, * Contact author: Przemyslaw.Biecek@pl.ibm.com

Keywords: IBM Netezza 1000, Netezza Performance Server, High Performance Cluster, parallel data processing

IBM Netezza 1000 is a warehouse that allows for using R functions in parallel in database very close to data. However in some cases user might want to offload the database and send computations to HPC server rich in RAM. In this talk we are going to present the prototype solution that benefits from parallel database IBM Netezza 1000 and the HPC server. Constructed framework facilitates implementing and executing parallel processing jobs. This is a next step to move analytics closer to big data and extend capabilities of in-database INZA procedures.

The main advantages of this approach are: reducing IBM Netezza 1000s overload by running computation-ally heavy tasks outside IBM Netezza 1000; more operational memory and faster processors on the cluster; dedicated hardware (e.g. GPU cards), since the cluster can be composed of diverse nodes; ultra-fast parallel data transfer via a dedicated connection.



- **Use case: Advanced analytics for Value at Risk in Operational Risk.** Data describing operational losses is stored in IBM Netezza 1000. Each row describes a loss, with corresponding risk category and business line. Modeling is done in parallel for each cell in the risk matrix. Data is distributed in IBM Netezza 1000 over risk categories and business lines and processed in a group-by-group fashion on the cluster. Each cluster node models losses distribution of a separate cell from the risk matrix.
- **Use case: Advanced Credit Scoring.** Both training and apply steps of the scoring process may be performed on the proposed architecture. Data describing credit or loans applications is stored in IBM Netezza 1000. In the first step different classifiers are trained on the whole training set. Model fitting and parameters tuning may be performed in parallel on different nodes. In the second step selected classifier or set of classifiers is applied in the row-by-row mode for all rows from the test/target dataset.
- **Use case: Bootstrap enhanced hierarchical genes clustering.** Data describing gene expressions is stored in the IBM Netezza 1000. The hierarchical clustering needs to be performed for bootstrap samples of the whole dataset. Thus the whole data set is distributed over computational nodes and the bootstrap operations are performed in parallel. The chunks of the dataset are sent to all computational nodes that broadcast their shares to all other nodes. Eventually, every node gathers the whole dataset. Bootstrap samples are generated and hierarchical clustering is executed on each node. Due to the high number of genes, each computational unit needs to be equipped with dozens of gigabytes of RAM. In the R statistical package different algorithms for hierarchical clustering are available, like fastcluster, which is order of magnitude times faster than standard implementation. Also, libraries that perform hierarchical clustering on GPU cards like gputools are available.

Slicing and dicing big data with RHadoop/rmr.

Antonio Piccolboni^{1,*}

1. Revolution Analytics

*Contact author: antonio@piccolboni.info

Keywords: Hadoop, RHadoop, **rmr**, mapreduce, big data

Hadoop has become the de-facto standard for storing and processing extremely large data sets. Some predict it will host half of all the world's data by 2016. Part of Hadoop is a parallel distributed computational model implementation known as Hadoop MapReduce whose abstract definition has roots in functional programming languages. As such, MapReduce is a natural fit for *R* where the `lapply-tapply` pair represents the closest analog. The package **rmr**, part of the open source RHadoop[1] project, provides an abstraction over Hadoop MapReduce that is tightly integrated with the R language but is capable of processing terabytes of data, taking advantage of clusters of up to thousands of machines. A simple library that works with most other R packages and in any IDE, **rmr** provides straightforward bridges between the in-memory and on-disk data, promoting a pragmatic and incremental approach to big data. Moreover, **rmr** supports the use of any *R* objects in connection with MapReduce, upholds for the most part usual *R* variable scoping rules, and doesn't force you to use any esoteric *R* constructs. This makes **rmr**-based programs look and feel and work like regular *R* programs providing what we believe is the easiest and most productive path into Hadoop and big data for *R* users and developers.

In this talk, after introducing Hadoop and the RHadoop project, we will illustrate, by way of simple examples, how to sift through and summarize large data sets with a few lines of code and how to go from simple one-liners to reusable functions.

References

- [1] Revolution Analytics (2011). The RHadoop project,
<http://github.com/RevolutionAnalytics/RHadoop>.

High Scale In-Database Modeling in Greenplum with R

Woo Jae Jung^{1*} & Noah Zimmerman^{1*}

1. Greenplum, a Division of EMC

*Contact authors: woo.jung@emc.com, noah.zimmerman@emc.com

Keywords: Greenplum, Revolution R, PL/R, in-database, MADlib

In a traditional analytics workflow using *R*, data are loaded from a data source, modeled or visualized, and the model scoring results are pushed back to the data source. Such an approach works well when (i) the amount of data can be loaded into memory, and (ii) the transfer of large amounts of data is inexpensive/fast. Here we explore the situation involving large data sets where these two assumptions are violated. In addition to moving large amounts of data to the computation in the modeling environment, we explore and demonstrate the paradigm of moving the relatively small computational components closer to the data, which is often referred to as ‘in-database’ or ‘alongside-database’ analytics [1].

The Greenplum database, a massively parallelized implementation of the popular PostgreSQL database, offers several alternatives to interact with *R* using the in-database analytics paradigm in a distributed environment. Here we explore four approaches to analytics in the Greenplum database: (i) procedural language calls to *R* from *SQL* (PL/R), (ii) reading data directly into *R* via standard DBI interfaces, and (iii) RevolutionR/Greenplum integration. Finally, we compare these results to the open-source analytics library MADlib [2].

We demonstrate the speed with which traditional analytics algorithms, such as logistic regression and k-means clustering, can be parallelized and run in-database. Through examples using data sources such as the 2010 US Census, we demonstrate situations where *R* analytics with Greenplum excel, and how advanced algorithms can be readily incorporated into a parallelized and highly scalable analytics workflow.

References

- [1] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Mad skills: New analysis practices for big data,” *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1481–1492, 2009.
- [2] <http://doc.madlib.net/>

penalizedSVM: feature selection for SVM classification in high dimensions

Natalia Becker^{1,*} and Axel Benner¹

1. Biostatistics, German Cancer Research Center, Heidelberg, Germany

*Contact author: natalia.becker@dkfz.de

Keywords: classification, SVM, feature selection, penalty function, high-dimensional data

Classification and feature selection play an important role in knowledge discovery in high-dimensional data. Support Vector Machine (SVM) algorithm is the most powerful classification and prediction methods with a wide range of scientific applications. The first implementation of SVM in *R* was introduced in the **e1071** package [1].

However, the SVM does not include automatic feature selection. The *R* package **penalizedSVM** [2,3] fills this gap. Regularization approaches extend SVM to a feature selection method in a flexible way using penalty functions like LASSO, SCAD and Elastic Net [4-6]. These functions are available in the package as wrappers, especially the LASSO SVM and the SCAD SVM have not been implemented in *R* so far.

Additionally, we proposed a novel penalty function for SVM classification tasks, Elastic SCAD, a combination of SCAD and ridge penalties which overcomes the limitations of each penalty alone [7].

Since SVM models are extremely sensitive to the choice of tuning parameters, the search of optimal tuning parameters is a special issue of the **penalizedSVM** package. In addition to the fixed grid approach, we adopted an interval search algorithm [8], which in comparison to a fixed grid search finds rapidly the global optimal solution.

The classification workflow is flexible, each penalization function is written as a wrapper, which allows developers to add easily new methods. The proposed methodologies are illustrated on real high-dimensional microarray datasets.

References

- [1] Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2011). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. *R* package version 1.6. <http://CRAN.R-project.org/package=e1071>
- [2] Becker N, Werft W, Benner A (2010). penalizedSVM: Feature Selection SVM using penalty functions. *R* package version 1.1. <http://CRAN.R-project.org/package=penalizedSVM>
- [3] Becker N, Werft W, Toedt G, Lichter P, Benner A (2009) penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics* 25: 1711–1712
- [4] Fung G, Mangasarian OL (2004) A feature selection newton method for support vector machine, *Computational Optimization and Applications Journal* 28: 185–202
- [5] Zhang, H (2006) Gene selection using support vector machines with non-convex penalty, *Bioinformatics*, 22(1), 88-95
- [6] Wang L, Zhu J, Zou H (2006) The doubly regularized support vector machine. *Statistica Sinica* 16: 589–615
- [7] Becker N, Toedt G, Lichter P, Benner A (2011) Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinformatics* 2011, 12:138
- [8] Froehlich H, Zell A (2005) Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In: *Int. Joint Conf. Neural Networks*, 1431–1438

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

E. James Harner^{1,*}, Shengqiao Li², Donald A. Adjeroh³

1. Department of Statistics, West Virginia University

2. UPMC Health Plan

3. Lane Department of Computer Science and Electrical Engineering, West Virginia University

*Contact author: jharner@stat.wvu.edu

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

Random KNN (RKNN) is a novel generalization of traditional nearest-neighbor modeling. Random KNN consists of an ensemble of base k-nearest neighbor models, each constructed from a random subset of the input variables. Random KNN can be used to select important features using the RKNN-FS algorithm. RKNN-FS is an innovative feature selection procedure for “small n, large p problems.” Empirical results on microarray data sets with thousands of variables and relatively few samples show that RKNN-FS is an effective feature selection approach for high-dimensional data. RKNN is similar to Random Forests (RF) in terms of classification accuracy without feature selection. However, RKNN provides much better classification accuracy than RF when each method incorporates a feature-selection step. RKNN is significantly more stable and robust than Random Forests for feature selection when the input data are noisy and/or unbalanced. Further, RKNN-FS is much faster than the Random Forests feature selection method (RF-FS), especially for large scale problems involving thousands of variables and/or multiple classes. Random KNN and feature selection algorithms are implemented in an R package **rknn**. We will show how to apply the Random KNN method via the **rknn** package to high-dimensional genomic data.

Reference

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

Sparse Principal Component Analysis using Stability Selection

Martin Sill^{1,*} and Axel Benner¹

1. Division of Biostatistics, DKFZ

*Contact author: m.sill@dkfz.de

Keywords: PCA, SVD, sparse, Stability Selection, high-dimensional

Principal component analysis (PCA) is a popular dimension reduction method. PCA approximates a numerical data matrix by a linear transformation of the original variables into a lower dimensional space that captures maximal variance of the data. Since each principal component is a linear combination of all variables of a data set, interpretation of the principal components can be difficult in case of high-dimensional data.

In order to find 'sparse' principal components that are linear combinations of only a subset of possibly relevant variables and therefore easier to interpret, several sparse PCA approaches have been proposed in the recent years. Typically, these methods relate the PCA to linear regression and perform a variable selection using penalty terms similar to those in penalized linear regression models.

We present a new approach to find sparse principal components of high-dimensional data extending sparse singular value decomposition (SSVD) [1]. Our approach combines regularized SVD with stability selection [2]. Stability selection is a general approach that combines variable selection with resampling techniques to control the error of falsely selecting irrelevant variables. Thus our new approach is able to find sparse principal components that are linear combinations of subsets of variables selected with respect to Type I error control. R code will be available in the package `s4vd`. Application of the method will be demonstrated using high-dimensional gene expression data sets.

References

- [1] Lee, M., H. Shen, J. Z. Huang, and J. S. Marron (2010, Dec). Biclustering via sparse singular value decomposition. *Biometrics* 66(4), 1087–1095.
- [2] Meinshausen, N. and P. Bühlmann (2010, July). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4), 417–473.
- [3] Sill, M., S. Kaiser, A. Benner, and A. Kopp-Schneider (2011, August). Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics* 27(15), 2089–2097.

Using the makeR Package for Managing Document Building and Versioning

Jason Bryer^{1,*}

1. University at Albany & Excelsior College

*Contact author: jason@bryer.org

Keywords: sweave, latex, building, versioning

R [3], \LaTeX [2], and Sweave [1] have proven to be incredibly useful for conducting reproducible research. However, managing document versions within R is limited. The **makeR** package attempts to provide the same ease-of-use for document versioning that the **devtools** [5] and **ProjectTemplate** [4] packages have provided for package development and data analysis, respectively. This package solves the problem where multiple versions of a document are required but the underlying analysis and typesetting code remains static or can be abstracted through the use of variables or properties. For example, many researchers conduct monthly, quarterly, and annual reports where the only difference from version-to-version, from an analysis and typesetting perspective, is the data input. Clearly R and \LaTeX are an ideal solution to this problem. The **makeR** package provides a framework to automate the process of generating new documents from a single source repository. This talk will cover the use of the **makeR** package for automating the document build process. Moreover, examples of extending **makeR** for project types other than Sweave and \LaTeX will be provided.

References

- [1] Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), *Compstat 2002 — Proceedings in Computational Statistics*, pp. 575–580. Physika Verlag, Heidelberg, Germany. ISBN 3-7908-1517-9.
- [2] Mittelbach, F. and C. Rowley (1999). *The \LaTeX 3 project*.
- [3] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- [4] White, J. M. (2011). *ProjectTemplate: Automates the creation of new statistical analysis projects*. R package version 0.3-5.
- [5] Wickham, H. (2011). *devtools: Tools to make developing R code easier*. R package version 0.4.

Rmarkup – a very simple tool for literate programming

Søren Højsgaard^{1,*}

1. Department of Mathematical Sciences, Aalborg University, Denmark *Contact author: sorenh@math.aau.dk

Keywords: markup, script file

The `Rmarkup` function in the `doBy` package does the following: Descriptive text and *R* code is put into a source document. The target document created by weaving will contain the descriptive text and program code together with graphics and results from the computations.

The format of the source document is a plain text file containing *R* and descriptive text (in lines starting with `##`). `Rmarkup` allows some markup facilities for the text. These are inspired by `txt2tags` markups (see <http://txt2tags.org/>). For example, **boldface**; *italics*; underline and `monospace` is produced with `**boldface**`; `//italics//`; `__underline__` and `&&monospace&&`. Moreover, different levels of headings are produced with `= Title level 1 =`; `== Title level 2 ==`; and so on. The target document is an HTML document containing the descriptive text (with possible markups), and program code together with graphics and results from the computations. `Rmarkup` is implemented by using the `RweaveHTML` driver in the **R2HTML**, [2].

A natural question is what `Rmarkup` offers that can not be accomplished using tools like `Sweave` [3] and `odfWeave` [1]. In terms of functionality, `Rmarkup` is nowhere nearly as advanced as `Sweave` and `odfWeave`. In terms of simplicity of installation and use, `Rmarkup` has advantages: `Rmarkup` grew out of teaching *R* to graduate students and others in the life sciences, e.g. students with a background in agronomy, biology, food science, veterinary science etc. Such students typically use Microsoft Office in their work and such students are generally hesitant to having to install too much software on their computer. We have found that requesting the students to install *R* itself and possibly also a suitable editor (we have recommended `Notepad++` for Windows users) is about as much as we can ask of the students. (Needless to say that it is pointless to ask such students to learn \LaTeX .) `Rmarkup` provides a tool for such *R* users.

We have found an additional value of `Rmarkup`: It is common to have a plain text file as a sandbox for playing around with e.g. data manipulation tasks. One may subsequently choose to form a \LaTeX file with these data manipulation steps and a textual description of steps using `Sweave`. In practice this means that one has a script file and a \LaTeX file describing essentially the same tasks, and hence there is a risk of things being out of sync. Using `Rmarkup` in connection with e.g. data manipulation tasks, there is only a simple script file in play and it is a manageable task to add useful comments and some simple text markup to such a file. This means that the task of documenting the work (in particular the more tedious parts of the work) is more likely to be done.

References

- [1] Kuhn, M. (2011). *odfWeave: Sweave processing of Open Document Format (ODF) files*. R package version 0.7.17.
- [2] Lecoutre, E. (2003, December). The R2HTML package. *R News* 3(3), 33–36.
- [3] Leisch, F. (2002). `Sweave`: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), *Compstat 2002 — Proceedings in Computational Statistics*, pp. 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

harvestr: A Simple Reproducible Parallel Simulation Framework

Andrew Redd

3/9/2012

Abstract

In the world of statistics and computing reproducibility is fundamental to research. In simulation studies this is complicated due to the perception that execution time is often considered a competitor to reproducibility. This is due to the fact that simulation can be greatly sped up through parallel computing techniques. Reproducibility in parallel situations is well researched but the tools are often cumbersome to use. I introduce the R package `harvestr`, which abstracts the concepts of parallel simulation to make them easier to program and use. Similar to the `plyr` and `foreach` packages which facilitate easy parallelization of code, `harvestr` facilitates the reproducibility in parallel simulations. This package facilitates the creation of simulations that are reproducible and fault tolerant.

Implementation of the Binomial Method of Option Pricing using Parallel Computing in R

Sai K. Popuri^{*}, Andrew M. Raim, Nagaraj K. Neerchal, Matthias K. Gobbert

Department of Mathematics and Statistics, High Performance Computing Facility (HPCF) and Center for Interdisciplinary Research and Consulting (CIRC), University of Maryland, Baltimore County

^{*}Contact author: saikul@umbc.edu

Keywords: RMPI, Options, Calls, Puts, Binomial model, Parallel computing

We explore the parallel computing capabilities in *R* with the **RMPI** package using the well-known Binomial method of Option pricing. We choose this problem because it naturally lends itself to a parallel computing framework by discretizing a continuous model in a lattice and using a backward induction method. This model is numerically intensive and demands a non-trivial communication protocol among individual nodes implementing the backward induction in parallel. We propose a general method to dynamically determine the number of nodes required as the problem size decreases. The root-node delegates the work to worker-nodes and lets the worker-nodes determine among themselves how to share and solve the problem. We believe this methodology is applicable to a wide range of lattice-based models where the domain space (in this case, time domain) is discretized and at each step, a business logic is applied to refine the computation before moving to the next step. We compare the performance of the Binomial model among three variations: a. traditional approach in serial mode, b. matrix approach in serial mode and c. traditional approach in parallel mode. We expect the traditional approach in serial mode to perform better for small number of steps but the parallel approach to perform better as the lattice becomes finer. As in any lattice-based model, the accuracy of the final result is linear in the number of the steps used to discretize. We investigate the effort spent on inter-node communication and the dynamic trade-off between computing and communicating. We increase the complexity of the problem to lay more emphasis on demonstrating the parallel computing capability in *R*. Although we have tested the implementation on the cluster tara in the UMBC High Performance Computing Facility, which has 82 computing nodes with two quad-core Intel Nehalem processors and 24 GB of memory, we believe the method can be ported on multi-core PCs and laptops.

Efficiently Executing *R* Code on Modern Hardware

Justin Talbot*, Zach DeVito, Pat Hanrahan

Stanford University, Department of Computer Science

*Contact author: jtalbot@stanford.edu

Keywords: High-performance computing, Runtime design, Just-in-time compilation

Recent hardware trends suggest that within the next couple years, the average data analyst's laptop will have an incredible amount of parallel hardware available—8 or more cores, each supporting hardware vectorized operations on 4 or more double-precision floating point values at a time (e.g. Intel's SSE and AVX instruction sets or the experimental Larrabee hardware). At present *R* has only limited capabilities for taking advantage of this parallelism. For example, the **parallel** package, recently added to the standard distribution, provides the ability to execute code on multiple cores simultaneously; but it requires the user to explicitly parallelize their code, and, due to implementation overhead, is limited to relatively coarse-grained parallelism. Additionally, the current *R* execution model is not well matched to the performance details of coming hardware. For example, most well-written (vectorized) *R* code, as executed today, is bottlenecked by memory accesses and will see little or no performance improvements as hardware vector sizes increase.

Over the last two years we have developed *Riposte*, a new research runtime for the *R* language, which has allowed us to experiment with a wide range of novel performance designs. In its current form, *Riposte* uses a dual virtual machine design inspired by previous work in high-performance APL implementations. The first virtual machine, a carefully implemented threaded interpreter efficiently executes scalar and short vector *R* code. While this runs, a low-overhead tracing method dynamically extracts any long vector code which is run by a second virtual machine. This second VM uses an optimization technique called *vector fusion* to eliminate most memory accesses, removing the memory bottleneck and permitting us to profitably use hardware vector units. The VM then does just-in-time compilation and automatically executes the code across multiple cores. The result is a high-performance, parallel *R* runtime that efficiently executes standard *R* code without any manual parallelization by the user.

At a level relevant for performance-oriented *R* users, we will discuss current trends in hardware technology and performance constraints in the current *R* implementation. Building upon our experience developing *Riposte*, we will then describe possible paths toward a high-performance *R*.

Building Distributed Intelligent Systems for the Enterprise with R

SMS Chauhan^{1,*}, Zubin Dowlaty¹

1. Mu Sigma Inc.

*Contact authors: smschauhan@gmail.com, zubin@dowlaty.com

Keywords: Enterprise messaging, Rules engine, R cluster, Intelligent Systems, Real Time Analytics, Software Agents

Distributed Intelligent Systems are a closely knit set of high performance components, each of which in itself has the capability to make a certain decision usually in low latency situations. Such systems are designed to scale to the needs of an enterprise as their usual implementations are in operational settings. The approach presented will discuss the essential building blocks of such systems and where R plugs in. We will introduce two packages, **Rjms** [1] and **Rdrools** [2] that supports the agenda of implementing R components in such systems. **Rjms** enables R to communicate with an enterprise message system such as Apache ActiveMQ [3]. **Rdrools** is a rules engine integration package based on Drools Expert [4]. Rdrools is utilized to apply transformations on a set of inputs based on predefined collection of rules. The capability to transform data along with the capability to send messages forms a powerful paradigm where R can be utilized in the formation of intelligent services. We will further discuss real world applications and best practices.

References

- [1] Rjms – R messaging using ActiveMQ and JMS, <http://cran.r-project.org/web/packages/Rjms>
- [2] Rdrools – A rules engine for R based on Drools. <http://cran.r-project.org/web/packages/Rdrools/>
- [3] Apache ActiveMQ– The open source messaging broker. <http://activemq.apache.org/>
- [4] Drools Expert– A rules engine commonly utilized in Java. <http://www.jboss.org/drools/drools-expert.html>

A Tale of Two Packages

Tim Hesterberg

Google, Inc.

TimHesterberg@gmail.com

Keywords: `dataframe`, `aggregate`

I'll talk about two packages. The **dataframe** package contains replacements for the data frame creation and subscripting functions in R, that dramatically cut the number of copies of that *R* makes of the data, e.g. from four to one for `as.data.frame(a vector)`, and from 10 to 3 for `data.frame` (a list). Naturally, this is faster. I'll give tips on how to improve your own code.

The **aggregate** package contains aggregation utilities, including improved versions of existing *R* functions, and new faster functions that drop down to *C* for speed.

cRy: a specialised software for statistical applications in x-ray crystallography

James Foadi^{1,*}, Gwyndaf Evans², David Waterman³

1. MPL, Imperial College, London SW7 2AZ

2. Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE

3. CCP4 - Research Complex at Harwell (RCaH) - Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA

*Contact author: j.foadi@imperial.ac.uk

Keywords: crystallography

A new package for applications in structural crystallography, **cRy**, has been developed in the renowned *R* statistical platform. This is the first software of its kind and it is supposed to provide a bridge between the large communities of crystallographers and professional statisticians. At present crystallographers make heavy use of large systems of programs, mainly written in Fortran, C/C++ or Python. These programs handle the several statistical operations, normally carried out in crystallography, with *ad hoc* routines, usually developed by different authors, often not sharing a common statistical platform. Data and results are exchanged through files with well-defined formats. The **cRy** package reads and writes files in the most commonly used crystallographic formats, carries out all major crystallographic calculations and provides an interface between the crystallographic data structure and the statistical objects and tools offered by *R*. **cRy** provides, thus, that common statistical platform that, at present, is still lacking in structural crystallography. The code has been developed using S4 classes.

cRy was presented at last year UseR conference in Warwick (UK) [1] and at the IUCr conference in Madrid [2]. Although it has not yet been completed in its final R-package form, interest for the software, and more generally for R, has been growing within the crystallographic community. The inclusion of a session on statistical packages at this year ECM in Norway [3] is a clear sign that packages like **cRy** will be more and more used in the coming years.

References

[1] UseR! 2011 Conference. *Coventry - UK*, 16-18 Aug (2011)

[2] International Union of Crystallography. XXII Congress and General Assembly *Madrid - Spain*, 22-30 Aug (2011)

[3] The 27th European Crystallographic Meeting *Bergen - Norway*, 6-11 Aug (2012)

crackR: Probabilistic Prediction of Airframe Fatigue Damage

Keith Halbert^{1,2,*}

1. The Boeing Company, Ridley Park, PA
2. Temple University Department of Statistics, Philadelphia, PA

*Contact author: keith.a.halbert@gmail.com

Keywords: reliability, aerospace, fatigue, damage tolerance

Fatigue, in metallic aircraft structure, is damage that occurs due to cyclic loading. Airframes are designed to continue functioning in the presence of unobserved small cracks. Inspections are scheduled at a frequency which is intended to lead to crack detection before cracks reach a critical size. The critical size is determined by either maximum repairable size or the fracture toughness of the material, the latter resulting in failure of the component.

The traditional methodology[1] for determining the inspection frequency for a specific structural feature is deterministic in nature. A crack of constant size, relative to the capability of the inspection method, is assumed to be present at manufacture. The growth of this crack can be reasonably estimated over time. By additionally assuming the material's fracture toughness is fixed, the engineer can determine the time until failure. Inspections are then scheduled at a frequency of one half of the estimated time until failure. While this has proven over time to maintain safety, there is no way to quantify the risk associated with the structure. Based on an assumed aggressive usage profile, maintenance plans determined in this way often call for far more inspections than necessary. In addition, the only way to predict the frequency of repairs is to examine similar structure on legacy aircraft. This leads to difficulties in accurately estimating fleet operating costs for new designs.

To estimate the risk of the structure and the likelihood of future crack detections, a probabilistic approach is required. This involves the use of random variables to represent the initial crack size, material fracture toughness, and various other values previously assumed to be fixed. Such an approach allows the user to find an optimal maintenance plan and evaluate alternative crack detection technologies[2]. The need for such technology has long been recognized within the aerospace community[3]. However, available software is closed-source, proprietary, or known to have stability issues; factors which have contributed to a lack of utilization of this methodology.

We present the recently developed R package **crackR**. This software is capable of calculating the probability of failure over the life of aircraft structure and the probability of finding cracks at future inspections. This package is intended to provide industry, government, and aerospace student users with free and extensible software for developing airframe maintenance plans. At the time of this writing the software is in beta mode.

References

- [1] Anderson, T.L., (2004) Fracture Mechanics: Fundamentals and Applications, Third Edition. *CRC Press, (Boca Raton, FL, USA)*
- [2] Halbert, K., Fitzwater, L.M., et al (2012). Cost/Benefit Analysis for System-Level Integration of Non-Deterministic Analysis and Maintenance Technology. *American Institute of Aeronautics and Astronautics Non-Deterministic Approaches Conference (Honolulu, HI, USA)*
- [3] USAF (2005). MIL-STD-1530C General Guidelines For Aircraft Structural Integrity Program

Measuring excess energy from galaxy feedback with mathematical modelling in R

Alastair Sanderson*

School of Physics & Astronomy, University of Birmingham, UK

*Contact author: ajrs@star.sr.bham.ac.uk (<http://www.sr.bham.ac.uk/~ajrs>)

Keywords: Differential equations, Simulation models, Data visualization, Astrophysics

R is an excellent environment for numerical analysis, in which complex mathematical models can be developed and explored, with easy manipulation and visualization of the results. I will describe how this approach can be applied to measure the excess energy resulting from cosmic feedback, primarily caused by outbursts from super-massive black holes.

Clusters of galaxies contain large amounts of hot gas in an approximately spherical dark matter ‘halo’. In this state, thermal pressure balances gravity and the entropy of the gas is very sensitive to non-gravitational heating. The gas temperature and density (and hence entropy) can be directly measured using X-ray observations [e.g. 1] and used to calculate its binding energy. This can then be compared to a baseline reference model which is subject to purely gravitational physics, in order to determine the excess binding energy from *non*-gravitational galaxy feedback.

However, the effects of heating are confounded with radiative cooling, which also raises the entropy of the gas (in this case by cooling the lowest entropy gas to form stars). It is therefore necessary to explore the impact of cooling on the baseline model, by truncating the entropy distribution. For a given entropy distribution and gravitational potential, the variation of gas temperature and density with radius can be determined by solving a coupled differential equation [3], subject to a choice of boundary conditions, using the **deSolve** package [2] in R. I will describe the practical experience of implementing this method and exploring a grid of models using the **plyr** package, with subsequent visualizing of the results using **ggplot2**.

References

- [1] Sanderson, A. J. R. and T. J. Ponman (2010, February). X-ray modelling of galaxy cluster gas and mass profiles. *Monthly Notices of the Royal Astronomical Society* 402, 65–72.
- [2] Soetaert, K., T. Petzoldt, and R. W. Setzer (2010). Solving differential equations in r: Package desolve. *Journal of Statistical Software* 33(9), 1–25.
- [3] Voit, G. M., G. L. Bryan, M. L. Balogh, and R. G. Bower (2002, September). Modified Entropy Models for the Intracluster Medium. *The Astrophysical Journal* 576, 601–624.

Metropolis-Hastings MCMC goes parallel on a GPU: first experiences and results

Giuseppe Bruno*

1. Bank of Italy, Economic research department
*Contact author: giuseppe.bruno@bancaditalia.it

Keywords: CUDA, GPU, Markov chain Monte Carlo, Metropolis-Hastings.

The evolution of graphical cards into complete double precision computing platforms has introduced another competitive element in the race for topping the highest Floating point Operations per seconds (*flops*) figure. Nonetheless the choice of statistical and econometrics algorithms eligible for parallelization require a careful analysis. In the last 5 years we have witnessed a significant increase in the number of statistical and econometric applications leveraging the availability of powerful and effective (*Gflops*/\$ that is billion of floating point operations per dollar spent) Graphical Processor Units (GPU). Many papers and books providing a wide range of examples on GPU programming have been written by computer specialists. Though we have to recognize a rather scant number of empirical applications carried out by research scholars in their domain. For example, at the research dept of the Bank of Italy, we have about 500 people. A good fraction of these people hold a PhD in economics or statistics but they prefer employing end-user packages like *Matlab*, *R*, *SAS* and *STATA*. Most of them would like to take advantage of the blazing speed of these GPU cards, but they do not want to waste their time in new mind-boggling computer languages. In this paper, we explore *R*'s capabilities for improving simulation performance through GPU functions. A very broad taxonomy splits the statistical simulation in two flavors: Monte Carlo and Markov Chain Monte Carlo (MCMC). Monte Carlo simulation belongs to embarrassingly parallel procedures. Parallelization of these algorithms is straightforward. On the other hand, MCMC simulation is based on correlated sampling. These procedures are not naturally convertible into parallel code. Here we tackle the parallelization of a MCMC method dubbed Metropolis-Hastings. This method is very well known in the statistical literature. Performance comparisons between CPU only and CPU+GPU code are provided. GPU functions have been developed by following the sample code available in the **gputools**. This is an *R* package developed and recently released by Buckner et al. [1]. It provides a set of useful primitives for running statistical analysis on the GPU within the NVIDIA CUDA™ framework. The natural question arising employing these GPU functions is: does the inequality $GPU_{time} \ll CPU_{time}$ always hold? Though in the paper we address only the Metropolis-Hastings algorithm, we show essentially two things: 1) by leveraging simple GPU aware functions it is possible to parallelize a fraction of algorithms not belonging to the embarrassingly parallel class; 2) the bandwidth of the communication channel between the CPU and the GPU and other hardware and software features establish the thresholds that make it worthwhile porting the code on the GPU. Starting from the paper by Jacob et al. [2] who provide CUDA *Python* functions, we write *R* code for Independent Metropolis-Hastings for both the CPU and the GPU. We show how is possible to leverage the computing capabilities of a GPU in a block independent Metropolis-Hastings algorithm. Considering the possibility of launching even 500 parallel threads on the presently available GPU cards, we can foresee the huge potential in this field.

References

- [1] Buckner, J., M. Seligman, and J. Wilson (2011). A gputools package. <http://cran.r-project.org/>.
- [2] Jacob, P., C. Roberts, and M. Smith (2010). Using Parallel Computation to Improve Independent Metropolis-Hastings Based Estimation. *Journal Of Computational And Graphical Statistics* 20, 1–18.

BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments

Michel Lang^{1,*}, Bernd Bischl¹, Olaf Mersmann¹, Jörg Rahnenführer¹, Claus Weihs¹

¹. TU Dortmund University

*Contact author: lang@statistik.tu-dortmund.de

Keywords: high performance computing, batch computing, experimental validation, reproducibility

Empirical analysis of statistical algorithms often demands time-consuming experiments which are best performed on high performance computing clusters. While distributed computing environments provide immense computational power, they are not easy to use for the non-expert. We present two R packages which greatly simplify working in batch computing environments.

The package **BatchJobs** implements the basic objects and procedures to control a batch cluster from within R. It is structured around cluster versions of the well-known higher order functions Map/Reduce/Filter from functional programming. The second package, **BatchExperiments**, is tailored for the still very general scenario of analyzing arbitrary algorithms on problem instances. It extends **BatchJobs** by letting the user define an array of jobs of the kind “apply algorithm A on problem instance P and store results R”. It is possible to associate statistical designs with parameters of algorithms and problems and therefore systematically study their influence on the algorithm’s performance.

An important feature is that the state of computation is persistently available in a database. The user can query the status of jobs, submit subsets of the experiments to the batch system and in a later stage add or remove experiments. Further, if SQLite is used as an backend, all data required to reproduce the experimental study, including internally handled seeds, resides in a single directory. The abstraction mechanism used to separate source code from cluster specific code makes the study portable to other batch systems.

Both packages are written in such a way that they can be used in principle on any batch system and even on machines without a batch system but that are accessible by SSH.

References

- [1] Bischl, B., M. Lang, O. Mersmann, J. Rahnenführer, and C. Weihs (2012). BatchJobs and BatchExperiments. <http://code.google.com/p/batchjobs>.

Scaling R to Internet Scale Data

Karl Millar^{1,*}, David Chen¹, Ni Wang¹

1. Google Inc

*Contact author: kmillar@google.com

Keywords: R, MapReduce, large data, distributed computing, scalable analysis

Analyzing internet-scale data sets requires statistical software that scales to data sizes several orders of magnitude larger than R is currently capable of handling. Tools such as MapReduce and Hadoop are capable of scaling to such large data sizes but are impractical for statisticians to use for data analysis.

We will discuss the overall design and API of packages designed to work over Google's distributed computing architecture that help to address these issues.

The foundation of this work is a package that provides an R version of the FlumeJava library[1] for distributed computation. This package provides a higher-level abstraction on top of Google's MapReduce framework, providing distributed generic collection classes and simple functions for manipulating them, which are automatically converted into an optimized sequence of MapReduces.

Building on top of this functionality, Google is building additional packages that provide a convenient interface for working with large data sets, including distributed versions of some of R's data objects and common functions and parallelized statistical algorithms to analyze large, distributed data sets.

References

- [1] Chambers, C., A. Raniwala, F. Perry, S. Adams, R. Henry, R. Bradshaw, and N. Weizenbaum (2010). Flumejava: Easy, efficient data-parallel pipelines. *ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*.

angleR: calculate with angles given in degrees, minutes and seconds

Andrej Blejec^{1,2}

1. National Institute of Biology, Ljubljana, Slovenia

2. University of Ljubljana, Ljubljana, Slovenia

Keywords: Angle, Degree, Geodesy, Circular statistics, Trigonometry

In modern trigonometry angles are measured in radians. However, in many branches of human activities angles are traditionally measured in degrees, minutes and seconds (DMS). To enable use of angles measured in DMS system in areas such as navigation or geodesy, we introduce a new *R* class `angle` and some supporting functions.

The class `angle` is implemented in S3. A set of supporting methods for conversions between different measurement units, arithmetic functions for addition, subtraction, and scalar multiplication as well as printing in traditional format (`° ' "`) is available. Trigonometric functions for class `angle` accept angles in DMS.

We will present the functionality of the class `angle` with possible use in traditional areas that are close to geometry, navigation, and geodesy that express directional data in DMS. We will also discuss some programming issues that were encountered during the development of planned package **angleR**. For statistical analysis, the class `angle` can be connected with the package **circular**[1].

References

[1] Agostinelli, C., Lund, U (2011). R package `circular`: Circular Statistics (version 0.4-3)

PL/R: The Fusion of PostgreSQL and R

Joseph Conway^{1,2,*}

1. credativ, LLC
2. PostgreSQL Global Development Group

*Contact author: mail@joeconway.com

Keywords: PostgreSQL, Database, SQL

PL/R is a PostgreSQL extension that allows the use of *R* from within PostgreSQL for advanced analytics in a simple, efficient, and controlled manner. It has been available and actively maintained since January 2003. **PL/R** supports PostgreSQL 8.3 through 9.1, and works with all recent versions of *R*. It is used by thousands of people worldwide.

Recent trends in big data favor bringing the analytics closer to the data – meanwhile **PL/R** has been quietly providing this service for over 9 years! This presentation will introduce the audience to **PL/R**, discuss the pros and cons of this approach to data analysis, take them through basic usage, and finally illustrate its power with a few advanced examples.

R as sms-processing and report-generating tool in exit-poll

Alexander Matrunich^{1,*}

1. Institute of Regional Development, Pskov, Russia

*Contact author: alexander@matrunich.com

Keywords: sms, exit-poll

In March of 2012 in Russia the elections of President took place. We'd got an order to run an exit-poll during voting day in Pskov region (it's similar to West Virginia by territory and to North Dakota by population). There were 100 voting stations in the sample. Hourly an interviewer from every station sent a sms-report with data about turnout, votes frequencies (there were 5 candidates) and amount of voters who refused to participate in the exit-poll.

Sms-gate was used for receiving and sending short text messages. Incoming sms-messages were available as csv-file, sending a message was performed as http-request. *R* was used for sms-text parsing (package **stringr**), evaluating of sms-reports, monitoring of interviewers' activity, feedback with interviewers and supervisors, generating of html-report(package **brew**). MySQL server was used for storing all the data (package **RODBC**).

Every sms-report was tested for existence of voting station, amount of numbers, logical validity (sum of votes frequencies and refusal can't exceed turnout). In case of error a request for correction was sent to the interviewer. In case of repeated error or silence from interviewer an alert was sent to the supervisor.

MySQL server allowed simultaneous of supervisors to view and edit manually status of voting stations and report data, if required. Generated report was published at web-server.

Some ideas we plan to realize: checking for duplicate sms-reports, checking and warning if votes frequencies differ markedly from previous hours or forecast, possibility of reserve sms-gate using, more user-friendly and secure web-interface for supervisors, more attractive and informative web-report.

A toolkit for cross-validation: The *R* package **cvTools**

Andreas Alfons*

ORSTAT Research Center, Faculty of Business and Economics, KU Leuven

*Contact author: andreas.alfons@econ.kuleuven.be

Keywords: Cross-validation, Package development, Prediction performance, Regression models

Cross-validation is a widely used technique in applied statistics to estimate the prediction performance of regression models on a given data set. In addition, advanced regression methods often have tuning parameters, and cross-validation is a popular way to select the final model.

The idea of cross-validation is simple and easy to implement: split the data into several blocks, leave out one block for model estimation, and predict the values of the left-out block. Those predictions are then used to compute a certain prediction loss function. Even though the basic procedure is simple, some additional programming effort is necessary for more complex procedures such as repeated (double) cross-validation, or using cross-validation to select the optimal combination of tuning parameters. While many packages for computing regression models already offer functionality for cross-validation, different packages use different interfaces and the returned objects have a different structure. Furthermore, developers often copy and paste the basic code skeleton when implementing cross-validation for different models, which complicates maintaining the code.

The *R* package **cvTools** [1] tackles those problems. It is mainly aimed at package developers and offers functions to carry out all iterations of complex cross-validation procedures with one simple command, including performing cross-validation for all possible combinations of tuning parameters of a method. The programming effort of implementing cross-validation for a certain model is thus greatly reduced. A typical function for cross-validation based on package **cvTools** consists of the following two parts: first the data is extracted from the model, then only one or two more lines of code are necessary to perform cross-validation.

The main advantage of **cvTools** is that functions building upon the package provide a unified interface to cross-validation and the same object structure for the results, which benefits users. Cross-validation methods for objects returned by popular regression functions are thereby already included in **cvTools**. Another advantage for users is that several cross-validation objects can be combined into a larger object, which then reports the optimal model. Moreover, different plots are available in **cvTools** for visual inspection of the cross-validation results.

References

[1] Alfons, A. (2012). **cvTools**: *Cross-validation tools for regression models*. *R* package version 0.3.0.

Smooth Bootstrap Inference for Parametric Quantile Regression

Tatjana Kecojević^{1*}, Peter Foster²

1. University of Central Lancashire, UK

2. University of Manchester, UK

*Contact author: tkecojevic@uclan.ac.uk

Keywords: Quantile regression, Bootstrapping, Kernel smoothing, General linear model, Confidence interval

Assessing the accuracy of the τ^{th} ($\tau \in [0, 1]$) quantile parametric regression function estimate (see Koenker and Bassett [1]) requires valid and reliable procedures for estimating the asymptotic variance-covariance matrix of the estimated parameters. This covariance matrix depends on the reciprocal of the density function of the error (sparsity function) evaluated at the quantile of interest which, particularly for heteroscedastic non-iid cases, results in a complex and difficult estimation problem. It is well-known that the construction of confidence intervals based on the quantile regression estimator can be simplified by using a bootstrap.

To construct confidence intervals in quantile regression we propose an effective and easy to apply bootstrap method based on the idea of Silverman's [2] kernel smoothing approach. This proposed bootstrapping method requires the estimation of the conditional variance function of the fitted quantile.

After fitting the τ^{th} quantile function, we obtain the residuals, which are squared and centered to zero. Estimating the conditional mean function of the centered squared residuals gives the conditional variance function of the errors about the estimated τ^{th} quantile. Using an estimate of the conditional variance function allows the standardisation of the residuals which are then used in Silverman's [2] kernel smoothing bootstrapping procedure to make inferences about the parameters of the τ^{th} quantile function.

In this talk we will discuss a variety of approaches to estimate the conditional variance function. These will include the adaptation of GLMs as well as non-parametric regression based estimation. These different approaches have been assessed under various data structures simulated in *R* and compared to several existing methods computable in the **quantreg** package contributed by Roger Koenker. The simulation studies show good results in terms of coverage probability and the spread of the constructed parameters confidence intervals when compared with existing methods.

This methodology is also applicable to a wider class of regression models with heteroscedastic errors where the transformation to normality is difficult to achieve or maybe undesirable given a need to preserve the original data scale.

References

[1] Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica* 46(1), 33–50.

[2] Silverman, B. W. (1986). *Density Estimation*. New York: Chapman and Hall.

Outliers and structural breaks in structural time series

Giovanni Petris

University of Arkansas
GPetris@uark.edu

Keywords: State space models, Bayesian inference, Hierarchical models

Structural time series are used as models for series with a smoothly changing level and/or trend. Although extremely useful in practice, in many real applications we may observe level (or trend) jumps in addition to the assumed smooth variation that the model accounts for. We propose a Bayesian hierarchical model that allows for such abrupt changes at any time as well as for observational outliers in structural time series. The implementation in *R* is based on package **dlm** [1], a powerful and flexible environment for Bayesian and frequentist analysis of dynamic linear models.

References

- [1] Petris (2010). An *R* package for dynamic linear models. *Journal of Statistical Software* 36(12).
- [2] Petris, Petrone, and Campagnoli (2009). *Dynamic linear models with R*. Springer, New York.

R, Quo Vadis?

Bill Venables^{1,*}

1. CSIRO Division of Mathematics, Informatics & Statistics, Australia

*Contact author: Bill.Venables@CSIRO.au

Keywords: Programming, Data Analysis, Modelling, R

Language designers seem to regard *R* as an ugly, inefficient language and hence find its popularity mystifying. See, for example, [3, 4].

I will present four examples from my own work, two large data analyses problems in fisheries, and two more abstract programming examples. Hopefully these will be of intrinsic interest, but together they encapsulate why I think users find *R* so invaluable. My thesis is that good data analysis and modelling require the practitioner to engage *interactively* with data, and that at some level *programming* becomes essential to this. This is essentially the same message as that presented in Chambers [1, 2], and the same idea implicitly underlies [5]. The popularity of *R* is primarily due to the way it provides support for this activity, making near optimal trade-offs. This view is mostly consistent with Cook [3] but there are some important differences. (The claim that *R* is *necessarily* ugly is also disputed!)

Although *R* may be well suited to meet many contemporary data analysis problems, it will not remain so indefinitely. I do not attempt to answer the existential question posed in the title, but rather suggest it as one we should be thinking about, now. I will present some thoughts on a SWOT assessment for *R*, and suggest ways we might prepare for a graceful transition to whatever becomes the next phase. Such a new phase, or phases, will inevitably come as data analysis itself rapidly evolves in both scope and scale.

References

- [1] Chambers, J. M. (1998). *Programming with Data: A Guide to the S Language*. New York: Springer-Verlag.
- [2] Chambers, J. M. (2008). *Software for Data Analysis: Programming with R*. New York: Springer-Verlag.
- [3] Cook, J. D. (2012). Why and how people use *R*. <http://channel9.msdn.com/Events/Lang-NEXT/Lang-NEXT-2012/Why-and-How-People-Use-R>. See also <http://lambda-the-ultimate.org/node/4503> and <http://lambda-the-ultimate.org/node/4507>.
- [4] Morandat, F., B. Hill, L. Oswald, and J. Vitek (2012). Evaluating the design of the *R* language: Objects and functions for data analysis. <http://www.cs.purdue.edu/homes/jv/pubs/ecoop12.pdf>. An ECOOP 2012 paper.
- [5] Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer-Verlag.

AUTHOR INDEX

- Çetinkaya-Rundel, Mine, 37
- Alfons, Andreas, 130
Allaire, JJ, 13, 107
Anderson, Brooke, 36
Arnold, Taylor, 49
Aude, Jean-Christophe, 32
- Barve, Narayani, 96
Barve, Vijay, 83
Bates, Douglas, 5
Becker, Natalia, 111
Biecek, Przemyslaw, 108
Blejec, Andrej, 127
Breheny, Patrick, 57
Brodsky, Jae, 58
Bruno, Giuseppe, 124
Bryer, Jason, 114
- Cano, Emilio L., 69, 91
Chan, TszKin Julian, 59
Chang, Winston, 86
Charrad, Malika, 51
Cheng, Joe, 13
Chine, Karim, 20, 60, 84
Choi, Leena, 61
Connolly, Daniel, 28
Conway, Joseph, 128
Cook, Di, 21
Coombes, Kevin, 82
Covington, Kyle, 29
Crookston, Nicholas, 97
- Damico, Anthony, 62
Dazard, Jean-Eudes, 30
de Jonge, Edwin, 78
Dowlaty, Zubin, 119
- Eddelbuettel, Dirk, 4, 14
- Ferguson, Nicole, 40
Fianu, Emmanuel Senyo, 92
Foadi, James, 121
François, Romain, 4, 14
- Gillespie, Colin, 38
Griffith, Sandra, 54
Grolemund, Garrett, 87
- Halbert, Keith, 122
Hanson, Bryan, 44, 63, 64
Harner, Jim, 112
Harrell, Frank, 1
Heiberger, Richard, 10
Hesterberg, Tim, 120
Hoffmeister, Sebastian, 65
Hong, Seonghak, 45
Horner, Jeffrey, 7
Hornick, Mark, 46
Højsgaard, Søren, 115
- Jablonski, Kathleen, 41
Jeon, Heewon, 66
- Kahle, David, 33
Kaplan, Daniel, 80
Kasturi, Jyotsna, 55
Kecojevic, Tatjana, 131
Kim, Keehoon, 67
Kolb, Jan-Philipp, 100
Kotthaus, Helena, 68
Kovalchik, Stephanie, 42
Kowalczyk, Paul, 31
Kuhn, Max, 17, 50
- Lang, Michel, 125
Lau, Olivia, 11
Lees, Jonathan, 98
Levy, Drew, 90
Ligges, Uwe, 3
- Maechler, Martin, 10
Matloff, Norm, 106
Matrunich, Alexander, 129
Maynes, Spencer, 70
McCann, Patrick, 47
McCush, Jackie, 93
Millar, Karl, 126
Morgan, Martin, 16

Muenchen, Robert, 18, 89

Nutter, Benjamin, 22

Olson, Joe, 103

Ooms, Jeroen, 85

Oren, Rodger, 71

Osvald, Leo, 88

Paulson, Josh, 13

Petris, Giovanni, 132

Piccolboni, Antonio , 109

Polzehl, Jörg, 8

Popuri, Sai Kumar, 117

Porzak, Jim, 43

Radziwill, Nicole, 72

Redd, Andrew, 116

Rickert, Joseph, 48

Robison-Cox, James, 73

Rossouw, Ruan, 94

Roth, Thomas, 104

Rounds, Jeremiah, 6

Rowlingson, Barry, 19

Rundel, Colin, 99

Rutter, Michael, 23

Sanderson, Alastair, 123

Schuette, Paul, 52

Scott, Terri, 1

Seier, Edith, 39

Sendecki, Jocelyn, 56

Sill, Martin, 113

Smith, David, 25

Stone, Eric, 26

Tabelow, Karsten, 8

Talbot, Justin, 118

Therneau, Terry, 15

Thomas, Samuel, 95

Timm, Andrew, 74

Turner, Heather, 27

van der Loo, Mark, 102

Venables, Bill, 133

Vendettuoli, Marie , 34

Verzani, John, 105

Wang, Shubing, 53

Wang, Wei, 81

Whitcher, Brandon, 8, 79

Wickham, Hadley, 12, 24

Wieczorek, Jerzy, 101

Woo, Jung, 110

Wright, Kevin, 75

Xie, Yihui, 35, 107

Yergens, Dean, 76, 77