Why R? Conference

Wrocław, Poland, 2-5 July 2018

Abstract Book

WHY R?
FOUNDATION

Committees

Scientific Committee

Adolpho Alvarez
Małgorzata Bodgan
Tomasz Melcer

Organising Committee

Michał Burdukiewicz
Marcin Kosiński
Stefan Rödiger
Alicja Gosiewska
Aleksandra Grudziąż
Malte Grosser
Andrej - Nikolai Spiess
Przemysław Gagat
Joanna Szyda
Paweł Mackiewicz
Bartosz Sękiewicz
Przemysław Biecek
Piotr Sobczyk
Marta Karaś
Marcin Krzystanek
Marcin Łukaszewicz
Agnieszka Borsuk - De Moor
Jarosław Chilimoniuk
Michał Maj
Michał Kurtys

Contents

Part I
Keynotes

1 Forecasting streamflow using the HydroProg system developed in R

Tomasz Niedzielski

HydroProg | Head of the Department of Geoinformatics and Cartography at the University of Wroclaw

The objective of the talk is to present the performance of HydroProg, which is a real-time and fully-automated system for issuing early warnings against high flows and floods. The HydroProg system is developed in R and its main task is to integrate hydrometeorological gauging networks with numerous hydrologic models in order to compute rapid prognoses of streamflow. The integration in question enables the production of multimodel ensemble predictions which are based on real-time weighting of forecasts computed by hydrologic models that are unlike each other. Two implementations of HydroProg in southwestern Poland are discussed: for the upper Nysa Kłodzka river basin (mountainous case) and for the upper and middle Odra river basin (predominantly lowland case). The accuracy and skills of the system are provided, with the particular emphasis put on forecasting high flows. The external modules that can be integrated with HydroProg are also mentioned. The activities towards elaborating a complete version of the system, suitable for complementing the reliable hydrologic forecasts issued by national and international institutions, are also presented. The research is financed by the National Centre for Research and Development as well as the National Science Centre, Poland (projects no. TANGO1/267857/NCBR/2015, 2011/01/D/ST10/04171).ed in R

2 The age of automation: What does it mean for data scientists?

Daria Szmurło

Jagiellonian University | Expert in Data Science | Analytics Hub Leader at McKinsey and Company

---

3 Project evolution – from university to commerce

Agnieszka Suchwałko

Wroclaw University of Technology | Data Scientist and Associate at QuantUp

One PhD student (physicist) comes up with an idea and shares it with the second PhD student (computer scientist). The second PhD student develops the idea and together they create an innovative method. The method turns out to be competitive compared to the available ones. Finally, it is bought by a private investor who sets up the company and prepares the method for sale. Story like many others? Probably yes, but this one uses R to identify the bacterial species based on the diffractograms of the bacterial colonies from the very first idea to MVP, and final product.

---

4 Machine Learning in R: package mlr

Bernd Bischl

Ludwig-Maximilians-University of Munich | A professor for Computational Statistics at the Department of Statistics

---

5 A business view on predictive modeling: goals, assumptions, implementation

Artur Suchwałko

Wroclaw University of Technology | Ph.D. | Data Scientist | Owner | Instructor at QuantUp

What is easy? To build a predictive model. What is hard? To build it to work as it was intended to. I'll tell what bad can happen during building and implementing of a predictive model and how to avoid serious troubles.

---

6 New advances in text mining: exploring word embeddings

Maciej Eder

Pedagogical University of Cracow | Head of Polish Language Institute, Polish Academy of Sciences (PAN)

---

7 Simulation of dynamic models in R

Thomas Petzoldt

Dresden University of Technology, Institute of Hydrobiology | Simulation of dynamic models in R | Senior Scientist, Faculty of Environmental Sciences

The world around us and the subjects we are working with are in permanent motion. Such time-dependent systems are called "dynamical", they change, move, develop, break down or evolve. Dynamic models aim to simulate this movement in the computer, either with discrete time steps or with differential equations to get a continuous time series. Unfortunately, only few differential equation models can be integrated analytically, while the more complex require a numerical simulation software. Dynamical simulations produce plenty of data. Their analysis was indeed one of my strongest motivations to start with R, a few years before the first useR conference. Around 100 add-on packages existed at that time, one of them 'odesolve' [1] with the 'lsoda' differential equation solver. I was very excited, planned to use it for an upcoming workshop about ecological modeling, but then missed an urgently needed feature, just one week before the workshop. I looked in the source code and got in contact with the package developer, Woodrow Setzer. The problem was quickly solved and then R took over and replaced almost all of our modeling tools - and our productivity increased. In 2009, Karline Soetaert's first book appeared [2], a kind of book that I had been really waiting for. She brought R's dynamic modelers toolbox a tremendous step forward by implementing a full set of solvers, by making them more user-friendly, and by contributing great examples, companion packages and comprehensive documentation. Having been invited to contribute to this team [3] was a great experience and it is still a pleasure to me.

I will present essentials of R's dynamic modeling ecosystem that evolved since then:

- essentials of core package 'deSolve', its solvers and extensions that make them convenient
- the way, how to speed up ordinary and partial differential equation models by using matrix algebra and compiled C or Fortran code
- package 'rodeo' [4] for generating Fortran code on the fly from Excel tables
- a demonstration how to embed dynamic simulations in web-pages with 'shiny' [5]

The power and versatility of the approach will be demonstrated by small teaching examples and two more complex research applications.

References

[1] Setzer, W. (2001) The odesolve package. Solvers for ordinary differential equations. Package version 0.1-1

[2] Soetaert, K. and Hermann, P.M.J. (2009) A Practical Guide to Ecological Modelling. Using R as a Simulation Platform. Springer-Verlag.

[3] Soetaert, K., Petzoldt, T. and Setzer, W. (2010) Solving differential equations in R: Package deSolve. Journal of Statistical Software 33(9), 1-25 DOI 10.18637/jss.v033.i09

[4] Kneis, D., Petzoldt, T. and Berendonk, T. (2017) An R-package to boost fitness and life expectancy of environmental models. Environmental Modelling and Software 96, 123-127. DOI 10.1016/j.envsoft.2017.06.036
[5] Chang, W., Cheng, J., Allaire, JJ., Xie, Y. and McPherson, J. (2018) shiny: Web application framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny

8 Deep Learning with R using TensorFlow

Leon Eyrich Jessen

Technical University of Denmark | Immunoinformatics and Machine Learning, DTU Bioinformatics Employees

The Deep Learning library TensorFlow, was originally developed by researchers and engineers working on the Google Brain team within Google's Machine Intelligence Research organization. In 2015 Google decided to release TensorFlow as open source software for all to use. The paramount criteria for any technology to have an impact on our society is availability. With the release of TensorFlow we have seen a revolution in application of artificial intelligence within numerous areas. Early 2018 RStudio released a suite of R-packages for working with TensorFlow in R. Thereby, R-users can now tap into hitherto unavailable AI technology. Keras, a high-level TensorFlow API, enables fast experimentation - Being able to go from idea to result with the least possible delay is key to doing good research. In my research I apply deep learning using Keras for R to unravel the rules governing molecular interactions in the human immune system.

Part II
Sponsor sessions

9 Kruk SA

9.1 Data-driven indulgence

Piotr Zieliński

9.2 R in pRoduction at KRuk

Aleksandra Gruchot

9.3 Why R Is a Perfect Tool for the Debt Portfolio Valuation?

Grzegorz Chłapiński, Piotr Michalski

Part III
Workshops

10 DALEX: Descriptive mAchine Learning EXplanations. Tools for exploration, validation and explanation of complex machine learning models

Mateusz Staniak

University of Wrocław

Complex machine learning models are frequently used in predictive modeling. There are a lot of examples for random forest like or boosting like models in medicine, finance, agriculture etc.

In this workshop we will show why and how one would analyze the structure of the black-box model.

This will be a hands-on workshop with four parts. In each part there will be a short lecture and then time for practice and discussion. Find the description for each part below.

Introduction Here we will show what problems may arise from blind application of black-box models. Also we will show situations in which the understanding of a model structure leads to model improvements, model stability and larger trust in the model. During the hands-on part we will fit few complex models (like xgboost, randomForest) with the mlr package and discuss basic diagnostic tools for these models.

Conditional Explainers In this part we will introduce techniques for understanding of marginal/conditional response of a model given a one- two-variables. We will cover PDP (Partial Dependence Plots) and ICE (Individual Conditional Expectations) packages for continuous variables and MPP (Merging Path Plot from factorMerger package) for categorical variables.

Local Explainers In this part we will introduce techniques that explain key factors that drive single model predictions. This covers Break Down plots for linear models (lm / glm) and tree-based models (randomForestExplainer, xgboostExplainer) along with model agnostic approaches implemented in the live package (an extension of the LIME method).

Global Explainers In this part we will introduce tools for global analysis of the black-box model, like variable importance plots, interaction importance plots and tools for model diagnostic.

Packages that we will use include mlr (Bernd Bischl and others), DALEX (Przemysław Biecek), live (Staniak Mateusz, and Przemysław Biecek), FactorMerger(Sitko Agnieszka, and Przemyslaw Biecek), pdp (Greenwell, Brandon), ALEPlot (Apley, Dan).

11 Introduction to Deep Learning with Keras in R

Michał Maj

Appsilon Data Science

With the release of the R Keras package (https://keras.rstudio.com/) (by JJ Allaire and Francois Chollet) at the end of 2017 / beginning 2018 the topic of artificial neural networks and especially deep learning in R became red-hot within the R community.

In this workshop you will get answers for the following questions:

What are fully conected and convolutional neural networks? How to build a sequential model in Keras (keras_model_sequential() function)? How to compile and fit naural netwrks in Keras (compile() and fit() functions)? How to add regularization to neural networks (L1, L2, dropout)? How to save and load existing models ? How to perform data ingestion and augmentation using generators? How to use pre-trained models and perform fine-tuning ? How to use callbacks?

12 Jumping Rivers - Shiny Basics

Roman Popat

Jumping Rivers

A quick introduction to creating interactive visualisations of data using shiny. The workshop will first make sure everyone is familiar with rmarkdown and htmlwidgets for creating a document with nice visualisations. We will then extend this knowledge by examining the shiny package for creating input output bindings to interaction with our R data structures. We will cover the basics of input and output for a shiny application and then explore creating our own page layouts. By the end of the workshop participants should feel comfortable getting started with creating their own shiny applications.

Recap/intro to markdown: A quick introduction/refresher on rmarkdown for document styling Adding some interactive graphs through html widgets

Input widgets and render functions: Extend a markdown document to run using shiny Adding input controls Using inputs to render output tables and graphs

Page layouts using shiny and shiny dashboard: Shiny and shinydashboard allow more control over page layouts Creating a layout with input and output "slots"

---

13 Jumping Rivers - Advanced Shiny

Roman Popat

Jumping Rivers

This workshop would suit attendees who are already comfortable with creating shiny applications. We will explore how to add functionality to our app using javascript packages and code. No real javascript knowledge is required to get started if you are a confident R programmer but the session will contain examples with written javascript. We will then explore how one might deal with routines in a shiny application that take a long time to run, or how to provide a good experience for simultaneous users of your app. We will then explore creating a standalone web served API to our R code and integrate the use of it into a shiny application.

Adding functionality from javascript code: -An introduction to using a javascript package with a shiny application -The basics of passing javascript values to a shiny app as inputs

Futures and promises for long running code: -An introduction to the wonderful promises package by rstudio's joe cheng promise pipes -With some small changes to your app, stop long running tasks from blocking the main application

Create and integrate with an external API: -Plumber is a great R package for creating a REST API on your R code and functions, we will explore how to get up and running with serving our R functions as an API -Integrate our separate plumber API with our shiny app

---

14 Maps in R

Piotr Sobczyk

OLX

Creating spatial data visualization is one of the coolest elements in the R toolbox.

In the workshop I shall show you tips to follow, pitfalls to avoid and hacks that might be either one of them :) Starting from a basic plot function, we will cover ggplot2 and finish with R packages that use interactive javascript libraries. You will get familiar with the full process from finding the right data, its processing in R up to preparing final data to plot.

---

15 From RS data to knowledge – Remote Sensing in R

Bartłomiej Kraszewski

Forest Research Institute

Remote sensing data from different sensors is a rich source of information for studying the natural environment, natural phenomena and monitoring some extreme phenomena, i.e. floods. Analyses and products made on remote sensing data are often essential for supporting decision-making processes in cities, forests and agriculture. Analyses of RS data are carried out for large areas which amounts to the use of advanced tools for their processing, i.e. databases or programming languages. For this type of analyses the R language is used more and more often. Its tools inventory in this area is still growing. R packages can be used for data analysis, processing and visualization.

The workshop aim is to present R language packages that can be used to work with remote sensing data. During the course packages for GIS analysis (rgdal, rgeos, sf), raster data processing (raster) and ALS data processing (lidR) will be used. The possibility of mutual data integration will be presented in order to obtain new information for later analyses and modelling using machine learning. The entire workshop will be carried out as a simple project of remote sensing data analysis in the forest environment. During the workshop lecturers will put emphasis on the practical use of R packages, which they usually use in their daily work in large remote sensing projects (LIFE ForBioSensing and RemBioFor) carried out by the Forest Research Institute in Sękocin Stary.

Co-host of workshop: Agnieszka Kamińska from Forest Research Institute

16 iDash - Make your R slides awesome with xaringan

Mikołaj Olszewski

iDash

Preparing a slide deck with the results of your research seem to be quite straightforward. You produce all the plots and tables in R and just paste them into PowerPoint, right? Or you might have gone a bit further and already used RMarkdown with ioslides, Slidy or Beamer. Those technologies however have many drawbacks. Their default look is quite outdated and it's hard to customise it and make each slide look exactly as you wanted. This might be especially problematic in case of companies that needs to follow strict brand guidelines.

This hands-on workshop will introduce participants to a different package called xaringan that solves all the issues. It allows to customise each slide entirely to suit needs of the most demanding users. Since it also uses RMarkdown, it allows to produce not only eye-catching but also reproducible results. Moreover, it allows to preview your slides dynamically in RStudio making your work much easier. It's also relatively easy to export the slide deck (natively in HTML) to a pixel perfect PDF. Join us if you want to learn how make your next R slide deck awesome!

17 Constructing scales from survey questions

Tomasz Żółtak

Educational Research Institute (Warsaw, Poland)

Surveys often include sets of questions on the same subject, designed to create more general indicators of views, attitudes, knowledge or other characteristics of respondents. Such an indicators allow for synthesis of information, drawing more general conclusions and reduction of random measurement errors. As continuous variables, they are also easier to use in further analysis.

However, the use of survey questions often involves a number of problems:

- answers are given on scales that can't be treated as continuous (eg. a Likert scale)
- response to the questions may depend on the way in which they are worded, eg. respondents may react a little different to negative statements
- respondents may have different styles of answering questions, eg. some may prefer more extreme answers than the other
- in self-assessment questionnaires some respondents may be inclined to give untruthfully answers indicating a higher level of knowledge or skills

Workshop participants will learn how to use R to:

- create scales based on sets of categorical variables using Categorical Exploratory/Confirmatory Factor Analysis (CEFA / CCFA) and IRT models

- use models with bi-factor rotation to deal with different forms of asking questions
- correct for differences in a style of answering questions asked using a Likert scale
- use the possibility to correct self-assessment knowledge/skill indicators using fake items

During the workshop R packages ‘polycor’, ‘mirt’ and ‘laavan’ will be used along with the data from international surveys: ESS, PISA and PIAAC.

18 Analiza i wizualizacja danych w R!

Alicja Gosiewska, Aleksandra Grudziąż, Marta Sommer, Magda Młynarczyk

R-Ladies Warsaw

Interesujesz się analizą danych? Znasz podstawy programowania? Miałeś styczność z R i RStudio?

Jeśli na wszystkie pytania Twoja odpowiedź brzmi “tak!”, to zapraszamy Cię do udziału w naszych warsztatach z analizy i wizualizacji danych w R! W czasie warsztatów nauczysz się jak w szybki i czytelny sposób przetwarzać dane m.in. przy pomocy pakietu dplyr oraz jak prezentować je w formie wykresów (również interaktywnych!) używając bibliotek ggplot2 i plotly.

Warsztaty będą podzielone na cztery półtoragodzinne bloki:

- podstawowe przetwarzanie danych (m.in. dplyr)
- podstawowa wizualizacja danych (m.in. ggplot2)
- zaawansowane przetwarzanie danych (m.in. dplyr, tidyr)
- zaawansowana wizualizacja danych (m.in. ggplot2, plotly)

Part IV
R in Natural Sciences

19 strideter: R package for identifying individual's steps from sub-second level accelerometry data of walking

Marta Karaś

Johns Hopkins University, Bloomberg School of Public Health

Wearable accelerometers can provide objective high-density measurements of human physical activity through recording movement acceleration. Recent advances in technology and the decreasing cost of wearable devices led to an explosion in the popularity of wearable technology in health research. For example, quantifying gait parameters has become increasingly important for epidemiological and clinical studies. Due to complexity and volume of accelerometry data, automatic and unsupervised methods for precise walking segmentation are needed.

The R strideter package implements method we propose to precisely identify beginnings and ends of individual's steps from sub-second level accelerometry data of walking. The method employs the continuous dictionary learning framework to identify strides (two subsequent steps) from accelerometry data. Precisely, we define data-derived baseline patterns, which we name as movelets, representing a population-specific stride. Next, we perform two-step strides segmentation by combining pattern-recognition with a maxima-detection approach to precisely identify beginnings and ends of individual's strides.

We demonstrate the proposed method using accelerometry data collected during 450-meter outdoor walk of 32 study participants wearing accelerometers on a wrist, hip and both ankles. We validate the performance of the method and discuss individual-specific gait characteristics.

20 R in LiFE ForBioSensing remote sensing project

Bartłomiej Kraszewski

Laboratory of Geomatics, Forest Research Institute

The R language is a very popular tool for acquiring, processing and analysing remote sensing (RS) data. Many packages of this language have been developed for analysis of the forest environment but also many other tools, non-dedicated to remote sensing, are used in such projects. Within the Life+ ForBioSensing project „Comprehensive monitoring of stand dynamics in Białowieża Forest supported with remote sensing techniques", which is carried out by the Forest Research Institute, the R language is a basic tool for processing spatial data. The choice of such a method for data analysis is not accidental: in the R language there are many developed packages dedicated to processing of remote sensing and GIS analyses. At the same time, through a simple syntax R, it is easy to use by non-programmers.

During the project, from 2015 up to 2019 annual large collections of remote sensing data are acquired, incl.: airborne and terrestrial laser scanning, satellite imagery, hyperspectral imagery, aerial imagery and various terrestrial measurements. Such a large database of multi temporal RS data is acquired for the first time both in Poland and in the world. The number of data and processing requires a novel approach that integrates data from multiple sources. A part of the project is to create innovative tools based on integration of R, C++ and PostgreSQL/PostGIS database. The developed methods of data integration and their analysis are innovative on a global scale.

The tools and R language are used in the project to:

- process large remote sensing data sets
- develop new methods of improving and controlling products
- determinate features of trees and stands
- statistical analysis and modelling
- results visualization

Co-authors of presentation: Agnieszka Kamińska, Krzysztof Stereńczak

21 Application of mlr and ggplot for machine-learning and visualisation on multiple depended variables

Jaroslaw Jasiewicz

Adam Mickiewicz University

We want to present application of mlr package and ggplot grammar of graphics to build a complete learning and presentation system addressed for soil science. Modeling of soils properties based on the satellite images is a leading task in modern remote sensing. While modeling of a single attribute is rather straight-forward process proper modeling of multiple features using multiple learners and several data transformations leads to the combinatorial explosion of possible variants of solutions. Also, prediction of different soil properties gain its highest performance with different learners and sets of model hiper-parameters so searching for the optimal solution for each parameter may be a tedious and time-consuming process but, first of all, difficult to preset due to multidimensional nature of the results.

Our data set includes 150 points sampled in the area near Pokrzywno (Poznań) from two separate field crops. Sixteen chemical properties of soils (nutrients) determined in the wet laboratory. An eight spectral channels data acquired from WorldView-2 repository was used to build machine learning regression models between nutrients and radiometric response of soil surface. We used several popular learners (glm, r-Forest, Qubist, k-SVM, kKNN, PLS-R MARS and CR-splines). Input data were used as a raw data, transformed by contrast-stretch and Box-Cox universal transformation.

There are three problems in our research:

- how to treat data collected from two similar, yer separate sources
- how a transformation of the data affects the performance of different learners
- which learner is performing best on each nutrient? We noticed that different nutrients gain the best prediction for different learners and the difference is significant

Such approach raises several problems at the level of data processing and processing:

- how to run multiple learning/prediction/performance analysis processes in a single routine
- how to present such four-dimensional (learner-transformation-nutrient-crop field) results in a clear and human readable form?

To solved that problems applied mlr package due to universality, transparent proceeding, and easiness of batch processing, then the results were transformed with dplyr and we prepared set of ggplot routines to clearly visualize four-dimensional results in a form of single 2D plot. The simple analysis of the plot indicates best learners for given nutrient as well shows which method of transformation minimizes the impact of sampling on various arable fields (Box-Cox transformation)

22 Machine-learning and R for purposes of plastic surgery – classification and attractiveness evaluation of facial emotions

Lubomír Štěpánek

First Faculty of Medicine, Charles University and Faculty of Biomedical Engineering, Czech Technical University in Prague

Many current studies come to a conclusion that facial attractiveness perception is data-based and irrespective of the perceiver. However, the ways how to analyse associations between facial geometric image data and its visual impact always exceeded the power of classical statistical methods. In this study, we have applied machine-learning methods to identify geometric features of a face associated with an increase of facial attractiveness after undergoing rhinoplasty, a plastic surgery procedure for correcting the form and functions of a nose. Furthermore, current plastic surgery deals with aesthetic indications such as an improvement of the attractiveness of a smile or other facial emotions, hence we explored how accurate classification of faces into sets of facial emotions and their facial manifestations is, since categorization of human faces into somatotypes should take into consideration the fact that total face impression is also dependent on expressed facial emotion.

Both profile and portrait facial image data were collected for each patient, processed, landmarked and analysed using R language. Facial attractiveness was measured using Likert scale by a board of independent observers. Multivariate linear regression was performed to select predictors increasing facial attractiveness after undergoing rhinoplasty. The sets of used facial emotions and other facial manifestation originate from Ekman-Friesen FACS scale but was improved substantially. Bayesian naive classifiers using e1071 package, regression trees (CART) via tree and rpart packages and, finally, neural networks by neural net package were learned to allow assigning a new face image data into one of the facial emotions.

Enlargements of both a nasolabial and nasofrontal angle within rhinoplasty were determined as statistically significant predictors increasing facial attractiveness. Neural networks manifested the highest predictive accuracy of a new face categorization into facial emotions. Geometrical shape of a mouth, then eyebrows and finally eyes affect in descending order the (quality of) classified emotion, as was identified using decision trees.

We performed machine-learning analyses to point out which facial geometric features, based on large data evidence, affect facial attractiveness the most, and therefore should preferentially be treated within plastic surgeries. Additionally, the classification indicated new possible facial somatotypes based both on facial geometry and expressed emotions, and suggested which facial features determine the final assignment into one of the classes of emotions the most.

References:

[]Pavel Kasal, Patrik Fiala, Lubomír Štěpánek, et al. “Application of Image Analysis for Clinical Evaluation of Facial Structures”. In: Medsoft 2015 (2015), pp. 64–70. URL: http://www.creativeconnections.cz/medsoft/2015/Medsoft_2015_kasal.pdf

23 PCRedux: machine learning helper tool for sigmoid curves

Stefan Rödiger

BTU Cottbus - Senftenberg

There are numerous examples of data with a sigmoid ('S'-shaped) curves in data science. One example is amplification curve data from quantitative Polymerase chain reactions (qPCR). The qPCR is an indispensable technology in human diagnostics and forensics. From an amplification curve quantitative information can be determined which can be used to assess diseases.

There are software packages, which offer workflows and criteria to process the qPCR data. That includes the preprocessing of the raw data, the fitting of non-linear models on raw data, the calculation of quantification points, the computation of amplification efficiency, the relative gene expression analysis, normalization procedures and data management. However, there is no open source software package that contains classified data sets and provides biostatistical methods for machine learning on amplification curves.

The PCRedux package contains functions and classified amplification curves for machine learning and statistical analysis. In addition, the PCRedux package contains extensive labelled data sets of amplification curves from various qPCR devices and detection chemistries. The amplification curves were classified (negative, positive) by a human. For curve shape based classification, the tReem() funftion was developed. To analyze the amplification curves methods such as change-point analysis, regression analysis, autocorrelation analysis and model fitting have been integrated. The pcrfit_single() function calculate more than 45 features from the amplification curves. This is useful for creating models and predicting classes (e. g. negative, positive).

Additional functionality in the package includes:

- decision_modus(), which calculates the frequency of classes in a data set
- earlyreg(), which calculates features by a regression analysis in the background region
- head2tailratio(), which compares the ratio of the head and tail
- hookregNL() and hookreg(), which attempt to detect a hook effect in the amplification curve

- mblrr(), which performs local robust regressions analysis
- performeR(), which performance analysis (e.g., sensitivity, specificity, Cohen's kappa) for binary classification
- encu(), which enables high-throughput data processing

A prototype of the PCRedux web server for the analysis of amplification curves using the PCRedux package is available from http://www.smorfland.uni.wroc.pl/shiny/predPCR/. The web server uses the machine learning mlr package as interface to a large number of classification and regression techniques. The PCRedux can be used for the extraction of features and for machine learning on amplification curves. The PCRedux package is an add-on package (MIT license) for open source statistical computing language and environment R.

24 Method Comparisons with Applications in Biomarker Research and Development

Andre Beinrucker

Thermo Fisher Scientific

In many areas of science, such as Chemistry, Biology or Physics, two different quantitative measurements procedures need to be assessed and compared to each other. As a simple approach, one could plot the results of the two methods in a scatterplot and draw a linear (least squares) regression line through the cloud of points. We explain in this talk, why this approach is usually not appropriate and present alternatives, such as Passing-Bablok regression and weighted Deming regression. We show how these methods are implemented in the R-package mcr (by E. Manuilova, A. Schuetzenmeister, F. Model [1]). Further, we explain how we use this R-package in the context of Biomarker research and development and point out difficulties and common misinterpretations of method comparison results. Finally, we show how method comparisons can be performed online using a freely available Shiny app by Burak Bahar [2].
[1] https://cran.r-project.org/web/packages/mcr/index.html
[2] https://bahar.shinyapps.io/method_compare/

Part V
Lightning talks

25 RNA-seq sequence analysis with R/Bioconductor

Barbara Kosińska-Selbi

The Biostatistic Group, Department of Genetics, Faculty of Biology and Animal Science, Wrocław University of Environmental and Life Sciences Wrocław, Poland

Bioconductor (www.bioconductor.org) is a bioinformatics open software that provides R tools for the analysis of genomic data, including this from next-generation high-throughput sequencing methods. The aims of Bioconductor are to provide access to powerful tools for statistical and graphic methods that help visualize genomic data. The project promotes high quality documentation, repeatability of tests and a package installation system, which is partly independent from the standard R mechanism. The function provided by Bioconductor biocLite() is an overlay on install.packages(), ensuring the installation of the correct version of packages, compatible with the Bioconductor version. Based on the data posted on the official website, there are 1560 software packages and it has two releases each year. The latest version of Bioconductor is 3.7. Increasing advances in biotechnology lead to new types of data, and data sets are rapidly growing in size, resolution and diversity. RNA-seq is a next generation sequencing method that is used to quantify the amount of RNA in a biological sample at a given timepoint. This type of data can be read into R in different way, depending on the basic format. RNA-seq data stores information on transcriptome sequence and is distributed in a form of millions of reads, i.e. small files, each containing information of several nucleotide sequence. Bioconductor provides four basic packages for reading RNA-seq data easyRNASeq, ShortRead, Rsamtools and GenomicAlignments. An important step in analysing next generation sequencing data is filtering. This step is computed by easyRNASeq and allows e.g. to reject the reads that don't align to the reference genome, that have insufficient sequencing quality or that contain many unknown nucleotides.

The ShortRead package provides functionality for working with FASTQ files from high throughput sequence analysis and it allows also working with BAM files. Rsamtools package provides an interface to BAM files. BAM files are produced by samtools and other software, and represent a flexible format for storing 'short' reads aligned to reference genomes. GenomicAlignments package serves as the foundation for representing genomic alignments within the Bioconductor project. Another important step is the genomic annotation of the reads to relate their locations to genes and other genomic features, such as gene transcriptional start sites or binding sites of other factors, such as enhancers. For this step, Bioconductor provides a couple of packages, such as biomaRt, genomeInterval, GenomicFeatures. One of the important elements of RNA-seq data analysis is its meaningful visualization. The aim of this step is to present the data and results in the most informative way. Before any statistical analysis RNA-seq data have to be normalized, the purpose of this step is to identify and remove sources of systematic variation, other than differential expression. For this step one of the popular packs is marray. Bioconductor gives a wide option of packages that allow to test differential gene expression such as DESeq2. One of the important elements of RNA-seq data analysis is its meaningful visualization. The aim of this step is to present the data and results in the most informative way. To visualize the expression of genes, heatmaps, which allow to present gene expression levels for many samples, are increasingly being used. Bioconductor provides a special package called ComplexHeatmap. This package allows for a highly flexible way to arrange multiple heatmaps and supports self-defined annotation graphics. The availability of Bioconductor software allow for conducting an RNA-seq analysis for persons who are no professional programmers, but are biologists which can then concentrate on the proper interpretation of results.

26 What is new in data.table

Jan Gorecki

data.table R package is under active development for long time. I would like to briefly present what was achieved in recent releases. Some interesting points will include usage of openMP to parallelize C code. data.table is also dependency for mlr package.

27 How to prepare for a job in data science?

Małgorzata Schmidt

Pearson

In early 2017, I enrolled in an internship at a data science team at Pearson while completing my BA in Mathematics at the Adam Mickiewicz University (AMU) in Poznań. Since then I've continued my internship at the company but I also enrolled in an MA programme in data science at AMU. It's the first programme available at a Polish university that's aimed at educating the next generation of data scientists in order to meet the growing market demand.

I decided to give this talk because many colleagues asked me about which of the two paths can better prepare a person for a career in data science. If I already landed a long-term internship in a professional data science team where I learn daily on the job, then why would I need more structured learning in the academia? In the near future, will getting an MA in data science give you a competitive edge in the market? Should you stop wasting time on internships, hackathons, or Kaggle-like competitions, and study for two more years instead?

While there is no clear-cut answer to these questions, I'll share my experience about the advantages and disadvantages of each path. I hope this talk provide some guidance to those who are about to start their data science career.

28 Going down to South Park to make some tidy analysis

Patrik Drhlík

Technical University of Liberec

South Park is a famous American TV show that tells a story of four nine year old boys. It is widely known as being very satiric and that most of the characters use lots of naughty words. In this talk, I will present my results of a text analysis done mostly using the R tidytext package by Julia Silge and David Robinson. The main question that I will answer is: Who is the naugthiest chracter in the series? Even those people who know the TV show will be surprised by the results. I will also show a simple sentiment analysis or episode popularity based on IMDB ratings. Do you think that the naughtiest episodes are more popular? We will find out.

29 Using Scrum for data science projects

Emilia Pankowska

Pearson

This talk will show how you can implement Scrum for conducting data science projects. Scrum is a framework for solving complex problems. It makes it easier to coordinate such aspects of teamwork as designing incremental solutions while not losing the big picture, within-team and cross-team collaboration, information flow, etc.

While these days Scrum tends to be associated with software development, I'll show how we adopted it to our data science projects. Some of the Scrum methods may seem counter-intuitive. For example, there's a focus on incremental deliverables, even if they're just mockups. I'll focus on this and other elements of Scrum that we found challenging, so that you too can use it to improve the quality and comfort of your work.

30 RStudio Connect: push-button publishing from your RStudio IDE

Curtis Kephart

RStudio

- Connect is an on-premises, commercial product from RStudio.
- Publishing your Shiny apps, RMarkdown, APIs, and plots is made easy.
- Limit access to authenticated users.
- Automate production and distribution of your data science workflows.

Part VI
mlr ecosystem

31 mlrFDA: functional data analysis with mlr

Florian Pfisterer

Ludwig-Maximilians-University of Munich | Department of Statistics

32 Deep Learning in mlr

Janek Thomas

Ludwig-Maximilians-University of Munich | Department of Statistics

33 rlR: Deep reinforcement learning with R

Xudong Sun

Ludwig-Maximilians-University of Munich | Department of Statistics

34 Interpretable Machine Learning in R

Christoph Molnar

Ludwig-Maximilians-University of Munich | Department of Statistics

35 Multilabel Classification with mlr

Quay Au

Ludwig-Maximilians-University of Munich | Department of Statistics

Part VII
Data analysis

36 Marketing analytics the Shiny way

Adolfo Álvarez, Alina Tselinina

Analyx

37 Using online updating algorithms to predict speedway results

Dawid Kałędkowski

ClickMeeting.com

Sport or computer games rivalries need up-to-date estimation of players strength to suggest relevant opponent, specify risk or value bets. Online updating algorithms perform very well when data flows are continuous or sudden, saving computation capacity and processing time. Examining performance of speedway riders, several methods will be presented, explained and applied using R. You will also learn who was the best speedway rider last year, who is at the top now, and how riders abilities changes in time.
R packages: sport, Rcpp

References:

[]Mark E. Glickman (1999): Parameter estimation in large dynamic paired comparison experiments. Applied Statistics, 48:377-394. URL http://www.glicko.net/research/glicko.pdf

[]Mark E. GLickman (2001): Dynamic paired comparison models with stochastic variances, Journal of Applied Statistics, 28:673-689. URL http://www.glicko.net/research/dpcmsv.pdf

[]Ruby C. Weng and Chih-Jen Lin (2011): A Bayesian Approximation Method

for Online Ranking. Journal of Machine Learning Research,12:267-300. URL http://jmlr.csail.mit.edu/papers/volume12/weng11a/weng11a.pdf
[]William D. Penny and Stephen J. Roberts (1999): Dynamic Logistic Regression, Departament of Electrical and Electronic Engineering, Imperial College

38 auditor: an R package and methodology for validation of any statistical model

Alicja Gosiewska

MI2 DataLab, Warsaw University of Technology

Predictive modeling is the branch of statistics concerned with finding a model that best reflects the data-generating process. Lots of machine learning algorithms in this area have been developed and is still developing, therefore there are countless possible options to choose from and a lot of ways to do it. Predictions of a poorly fitted model will be misleading when confronted with future data, which is unacceptable and potentially hazardous in many cases, for example in medicine. That is why methods that support the selection of proper model are important.

During this talk, I will introduce the auditor package which is a uniform interface to statistics and visualizations that facilitate assessing and comparing the goodness of fit, performance, diagnostic and similarity of models. As it concentrates on the analysis of residuals, most of the presented methods are model-agnostic. In this talk I will present examples for classification and regression models. Moreover, the auditor is designed to make validation more convenient and accessible by reducing the amount of effort needed to create informative visualizations, for both commonly used plots and new approaches.

References:

[]Alicja Gosiewska and Przemyslaw Biecek (2018). auditor: Model Audit - Verification, Validation, and Error Analysis. R package version 0.2.1. [https://mi2-warsaw.github.io/auditor/]
[]Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. ""ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages."" The R Journal 8.2 (2016): 478-489

[]Przemyslaw Biecek (2018). DALEX: Descriptive mAchine Learning EXplanations. R package version 0.2.0. [https://CRAN.R-project.org/package=DALEX]

39 Artificial Intelligence in turn based strategy game

Łukasz Wawrowski

DOJI - Educational Innovations

Artificial Intelligence (AI) is one of the areas that use machine learning methods. It is mainly used in games, but more and more often appears in everyday life. As history shows, creating an engine of artificial intelligence that will defeat a human in a chess game or Go is not a simple task. Artificial intelligence in computer games can use a simple set of rules, tree search methods or even neural networks. During work breaks R can be used for entertainment. There is a fun package in which we can find some simple games. The goal of the presentation is to try to implement a strategic turn-based game in R similiar to the combat system in Heroes of Might and Magic. The game will use the alpha-beta algorithm, which will be an artificial intelligence of the computer.

40 Get your Machine Learning workflow with under control with DVC

Mikołaj Bogucki

Pearson

Data Version Control (DVC) is an experiment management tool that has been built to seamlessly integrates with the Git version control system. A DVC pipeline can support a data scientist in each step of the Machine Learning (ML) life cycle: from modifying ML algorithms (such as adding a new features or tuning hyperparameters) to ensuring reproducibility.

In this talk, I'd like to share my experience from using DVC in my work in the Advanced Computing and Data Science Lab at Pearson. I'll show how to build a DVC pipeline step by step. I'll also share my experience with building and managing the process of model development with DVC.

41 Correcting for SAMple BIAs with the sambia R package

Norbert Krautenbacher

Institute of Computational Biology, Helmholtz Center Munich | Chair of Mathematical Modeling of Biological Systems, Technical University Munich

Samples taken from a population can lead to sample selection bias, for instance because of complex survey designs, and can distort statistical analyses. We introduce the R-package 'sambia' which is a collection of various techniques correcting statistical models for sample selection bias. We focus on correcting arbitrary classifiers for biased samples resulting from stratified random sampling.

In particular, we place the resampling-based methods ""stochastic inverse-probability oversampling"" and ""parametric inverse-probability bagging"" at the disposal which generate synthetic observations in order to resemble the original data and covariance structure. These methods have been proposed and the latter shown to outperform state-of-the-art methods for random forests in Krautenbacher, Theis, and Fuchs, "Correcting Classifiers for Sample Selection Bias in Two-Phase Case-Control Studies", Computational and Mathematical Methods in Medicine, 2017.

Part VIII businessR

42 Dividing space - spatially constrained regionalization method for large vector GIS data

Adam Dąbrowski

Adam Mickiewicz University

Spatial segmentation or regionalization is a common problem in geography that in fact is an unsupervised spatially constrained clustering problem. Often, area under analysis is so complex it has to be divided into smaller subregions, that could be defined as internally homogenous, to perform further research. Such situation occurs while modeling real estate prices in a big city or when you are planning an experiment that requires stratified sampling. Current research concentrates on methods for segmenting raster data (usually satellite imaging or landcover data) however sometimes a need arises for understanding the structure of multidimensional data concerning diverse variables like demography, accessibility or visibility. Existing methods, in such situations, use common clustering methods like k-means or hierarchical clustering and incorporate spatial coordinates as complementary variables. Unfortunately, since the coordinates are only 2 of many variables the output regions often don't preserve spatial continuity. Other algorithms like SKATER (Spatial 'K'luster Analysis by Tree Edge Removal) are time consuming and don't work efficiently with very large datasets. I propose a new approach to such regionalization which uses neighborhood class as a topological structure from which a spatial network is being build. Basing on the similarity of adjacent objects, calculated from selected variables, they are being joined to their most similar community using network analysis. The proposed method combines spatial econometric and network analysis in an innovative way that expands the possibilities of spatial regionalization using

point or polygon GIS data but also enhances the possibility of analyzing hierarchical structure of the subregions.

43 Multi-state churn analysis with a subscription product

Marcin Kosiński

Gradient Metrics

Subscriptions are no longer just for newspapers. The consumer product landscape, particularly among e-commerce firms, includes a bevy of subscription-based business models. Internet and mobile phone subscriptions are now commonplace and joining the ranks are dietary supplements, meals, clothing, cosmetics and personal grooming products.

Standard metrics to diagnose a healthy consumer-brand relationship typically include customer purchase frequency and ultimately, retention of the customer demonstrated by regular purchases. If a brand notices that a customer isn't purchasing, it may consider targeting the customer with discount offers or deploying a tailored messaging campaign in the hope that the customer will return and not "churn".

The churn diagnosis, however, becomes more complicated for subscription-based products, many of which offer multiple delivery frequencies and the ability to pause a subscription. Brands with subscription-based products need to have some reliable measure of churn propensity so they can further isolate the factors that lead to churn and preemptively identify at-risk customers.

During the presentation I'll show how to analyze churn propensity for products with multiple states, such as different subscription cadences or a paused subscription. If the time allows I'll also present useful plots that provide deep insights during such modeling, that we have developed at Gradient Metrics - a quantitative marketing agency (http://gradientmetrics.com/).

44 Unnatural language processing

Yizhar Toren

Shopify

In the twilight zone between carefully managed keywords and natural language corpuses, we find a strange species of short text strings used by our suppliers to describe the products they upload into our catalog (aka product "titles"). These short phrases (5-30 words) contain a huge amount of information about the product (and therefore seemed a good place to start mapping our catalog) but do not follow a strict structure or natural language rules. In this talk I'll describe how I used R's data and NLP tools to handle the strange nature of these creatures at scale, and how I was able to quickly prototype a mechanism that serves our recommendation and search engine and helps us understand and improve our product catalog.

45 Tested on Humans

Mateusz Otmianowski

Pearson

As data scientists, we usually create data products such as dashboards not for our own use but to help others get their insights from data. A common mistake made by a beginner data scientist is to create a data product based on best data visualisation practices, and release it without testing it on target users. This often results in a product that places eye candy above usability.

This talk will show you how to can implement usability testing in your data science workflow. I'll walk you through the process of preparing for a usability test, running it, and analysing its results. I'll illustrate some of the design principles on the example of data products developed by the Advanced Computing and Data Science Lab at Pearson. I'll also share what worked and what didn't work for us along the way.

References:

[] Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems, Steve Krug

[] Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability, Steven Krug
[] Designing Interfaces, Jenifer Tidwell
[] Storytelling with Data: A Data Visualization Guide for Business Professionals, Cole Nussbaumer Knaflic

46 insuRvey: An R package for Instagram Based Surveys

Olgun Aydin

PhD. Candidate

As all of us know that most of the companies have started to make some surveys to make people pay attention to their Instagram pages and make them follow their Instagram pages. People can attend these surveys just with adding comments and mentioning their friends to survey related posts.

insuRvey library has been built to provide solution to Instagram Surveys. The library which has been built by using R, first of all navigates the url of survey related post, and collects comments, parses comments to detect users who put comments and users who are mentioned. At final step thanks to random sampling techniques, there is a chance to find the lucky person.

47 The R language on the Microsoft AI Platform

Bart Czernicki

Microsoft

Part IX
High peRfoRmance computing

48 Tricks for faster R code

Tomasz Melcer

QuantUp

Unless we are working with very small datasets, computation time is something that slows down analytics of any kind, whether business-related or academic. We usually need to run analyses many times, and if data processing takes more than a short while, human mind starts to wander off and we get less efficient. To waste less time on waiting for results, we may attempt to make our code faster. This is, however, a trade-off: optimizing code takes human time too! I will talk about some tricks that can speed up code with little effort: memoization and parallelization. I will show some common cases where these techniques bring most benefits, but also discuss when to avoid them.

49 Using the GPU for speeding up custom models in R

Krzysztof Jędrzejewski

Pearson

Estimating values of latent model parameters is one of the most common tasks in the work of a data scientist. For most popular models, there are specialised packages that can be used for that purpose. Yet these libraries are usually limited to the most generic forms of these models, and they do not allow for much customisation. Also, they are sometimes limited by the assumptions made by their authors.

One solution for such cases is to use libraries implementing dataflow programming paradigm such as TensorFlow. They're used most often for deep learning but they can be successfully applied to other tasks too. The added value of using such tools is the possibility to use GPU for processing, which can speed up computations up to 50 times. I'll talk about my journey into the world of IRT modelling. It started from using a specialised packages like TAM, but later led me to experimenting with a variety of other, more general tools. I'll show how Tensorflow, and other tools I tried, may be used to estimate model parameters when working with datasets containing millions of observations. I'll also compare algorithm performance results for both CPU and GPU and point out the advantages and disadvantages of using each technology.

In addition, I'll discuss how cloud services such as AWS can be used to build a highly scalable workflow for periodical recalculations of parameters of such models, even when the total amount of data exceeds a billion observations.

50 How to multiply your R computing power in 15 minutes for few bucks

Mikołaj Olszewski

iDash s.c.

R users are often running into performance troubles while trying to perform computationally intensive calculations on large datasets. Personal computers rarely have more than few cores and 16 gigabytes of RAM. While usually a lot of optimisation can be done to fit calculations into personal machine, sometimes it's just not enough.

One of the solutions is to buy professional hardware dedicated to perform calculations for you. This however, can be very costly and inefficient as you probably wouldn't use the available power all the time. Additionally, you're the one responsible for maintaining this machine, what simply cost some extra time if something breaks down.

The other option is to leverage the power of cloud computing. Setting up a powerful machine in the cloud, that can be used on demand, is now easier, quicker and cheaper than ever. In my presentation I'm going to live show how one can set up a fully functional machine running R and RStudio, having 64GB of ram and 8 cores in less than 15 minutes.

51 Tell me which packages you use and I'll tell you who you are (tidyverse vs data.table vs base R)

Anna Skrzydło, Bartosz Kowalski

MediaCom Business Science

Have you ever consciously considered which R packages you should use? If yes, then probably you've asked yourself such questions:

- Should I use just pure base R and never be dependent on package updates?
- Or maybe is it worth to switch my whole code to clear tidyverse environment?
- Oh wait, how about taking advantage of data.table efficiency?

With so many good options, the only right answer is – it depends... We've also spend some time experimenting with different combinations in our 15-person team and now have some insights to share with you.

Part X Web scraping and data visualisation

52 R as a tool for visualization of large demographic datasets

Anna Dmowska

Adam Mickiewicz University, Institute Geoecology and Geoinformation

The aim of this presentation is to provide a comprehensive framework for visualization of high resolution demographic data using dot density maps. The proposed algorithm to create dot density map is based on high resolution geospatial raster data provided as a result of dasymetric modeling.

Demographic data are usually visualize based on statistical data aggregated over previously defined regions by assigning one color to whole aggregated units (choropleth map) or by using dot density maps (e.g racial dot map). In such maps dots are randomly distributed within region and each dot represents one or more people. The main limitation of such maps is their dependence on the size of the region; uninhabited areas are not excluded from mapping.

The one solution used to produce more detailed maps is a dasymetric modeling which allow to redistribute demographic data (such as population or racial/ethnicity) from aggregated units into geospatial grids cells. The resultant map will be a high resolution geospatial raster dataset. Such map presents more detailed information about racial/population distribution by shifting all population only to inhabited areas and assigning it to grid cells using different weights for different types of residential areas. However calculating such maps for large areas is time consuming and it required efficient, fully automated, flexible procedure. So far there have been also no algorithms that would allow to use high resolution raster data to produce dot density maps.

We developed an automatic procedure which use R to build fully automated computational environment to work with demographic data in the continental scale (11 millions of records in tabular data and 8 billions of cells). This procedure consists of 2 steps: 1) performing dasymetric modeling to disaggregate population and race/ethnicity groups data into grid cells (this step was presented during ERUM 2016 conference); 2) using high resolution grids to visualize data as a dot density maps.

R provides us a tool to construct dot density maps based on statistical data aggregated in spatial units (maptools::dotsInPolys), but that algorithm flaws for serious limitations: it works only with large polygons and is very inefficient. Our solution uses raster data as an input and dots are randomly scattered in each cell. In situation when approximate number of people for one cell is below 1 algorithm uses probabilistic approach to decide whether to place a point in a given cell or not. If visualization cover more than one race it also build random stack where probability of displaying a point at the top depends on the percentage of the race in a cell.

53 Making Shiny shine brighter with 6 useful packages

Krystian Igras

LabMasters | MIM UW | Appsilon

There is no need to praise Shiny for its influence on presenting results. As with many other technology stacks, Shiny could benefit from community contributions for further development of the package itself and the growth of independent packages that add new features.

In this presentation we will show six packages that add interesting capabilities to Shiny such as beautiful UI shiny.semantic and semantic.dashboard, routing shiny.router, authentication shiny.users, app usage monitoring shiny.admin and internationalization shiny.i18n.

We will demo the app that uses all these packages. We show how their usage improves working with Shiny and what their development adds to the open source community. Presented packages help companies adopt R/shiny and proves to be very useful based on the feedback for out clients.

54 Webscraping with RSelenium

Pieter Krsteff-Jantcheff

Ordina Belgium

During the 3 years I have been working as a data scientist, it happened very rarely that I received a clean and tidy excel file with data to do analysis on. Very often all we have is a business question and we have to acquire the data from any source possible, like from the worldwide web. Therefore web scraping is a very powerful tool to have in your arsenal as a data geek. Despite the importance of it, I feel that this skill is underrated within the data science community, which is why I want to share my web scraping experience with you and introduce to you my favorite package in R to do this: RSelenium.

RSelenium makes it possible to connect to a Remote Selenium Server from R and provides R bindings for the Selenium Webdriver API. More specifically it allows you to extract any piece of information from a web page into R in a structured and automatic way. R users are easily scared away because it involves some Java syntax. This causes them to learn other web scraping packages. This is a shame because I believe RSelenium offers more freedom and flexibility than these other packages. After this talk you will know exactly what I mean.