eRum 2016
european R users meeting

October 12-14, 2016 in Poznan, Poland

Welcome Address

Welcome! Have a good time!

Contents

Invited Talks

Addressing the Gender Gap in the R Project

Heather Turner[1⋆]

1. University of Warwick

⋆Contact author: ht@heatherturner.net

Keywords: R community; package development

Despite **R**'s origins in the discipline of statistics and the strong uptake of **R** in fields such as ecology and genomics, where women are well represented in the workforce, the **R** developer community looks more like that of computer science generally, where women are in a distinct minority. If we consider contributions to the **R** project, the situation is even worse. How can we encourage more women to become developers and leaders in the **R** community?

Earlier this year the **R** Foundation, a not-for-profit organisation set up to support the **R** project (`https://www.r-project.org/foundation/`), established a task force to explore this question and to take actions to address the gender gap (`http://forwards.github.io/`; `https://twitter.com/RWomenTaskforce`). In this talk I will give an overview of the activities of the taskforce so far and our plans for the future. I will also share some tips for women looking to get more involved and ideas of ways everyone can help to make our community more inclusive.

Heteroscedastic Discriminant Analysis and its integration into 'mlR' package for uniform machine learning

Katarzyna Stąpor[1⋆]

1. Institute of Computer Science, Silesian Technical University

⋆Contact author: katarzyna.stapor@polsl.pl

Keywords: discriminant analysis; machine learning; heteroscedasticity
R packages: mlR

The *mlR* package (machine learning in **R**) offers a unified interface to access various machine learning

algorithms from other packages in **R**. This framework provides supervised methods like classification, regression and survival analysis along with their corresponding evaluation and optimization methods, as well as unsupervised methods like clustering. It is written in a way that you can extend it yourself or deviate from the implemented convenience methods and construct your own complex experiments or algorithms. As an example, it will be shown how to integrate into the *mlR* package the new, proposed by us learner, the Heteroscedastic Discriminant Analysis (HDA), being the extension of the classical Fisher Linear Discriminant Analysis (FDA), implemented in the *base* package. HDA is the extension of FDA for dealing with the case of unequal covariance matrices in the populations, the situation that occurs very often in practice. The new implemented permutation test for testing the equality of covariance matrices will be also presented. Integration the new learner into *mlR* requires defining the learner itself with the name, description, parameters, and a few other things, then providing the function that calls the learner function and builds the model given the data, and finally, a prediction function that returns predicted values given new data. The example of usage the new HDA learner on the real world credit dataset will also be presented.

A survey of tools for Bayesian data analysis in R

Rasmus Bååth[1*]

1. Lund University

*Contact author: rasmus.baath@gmail.com

Keywords: Bayesian data analysis; MCMC
R packages: rjags; rstan; MCMCpack; R-INLA

Bayesian data analysis is an intuitive and straightforward framework for doing both inferential statistics and predictive analysis. Because of its popularity there is now a plethora of **R** packages for fitting Bayesian models. This talk will guide you through the jungle and survey some of the most useful tools for doing Bayesian data analysis in **R**. The talk will feature concrete code examples throughout and does not assume that you have much prior knowledge about Bayesian statistics.

Browse Till You Die: Scalable Hierarchical Bayesian Modeling of cookie deletion

Jakub Glinka[1*]

1. GfK SE, Nuremberg

*Contact author: Jakub.Glinka@gfk.com

Keywords: hierarchical bayesian modeling; model based machine learning; consensus MCMC
R packages: rstan; parallelMCMCcombine; Rcpp

The common approach for tracking the device's on-line movement is through cookies - small portion of information stored within user Internet Browser. This enables Market Research companies to assign web browsing data to one specific browser. The main problem within cookie tracking framework is to assess

whether on given day the lack of its activity is due to the real absence or deletion. In usual site-centric approach one can only observe cookie's digital footprint on limited number of media providers which leads to the highly skewed data, moreover the deletion moment is not directly observable. In order to deal with mentioned challenges we designed Hierarchical Bayes model of the cookie behavior that enables us to pose questions about the probability of the cookie deletion. We will present how it fits Model Based Machine Learning Paradigm and how one can effciently estimate model coefficients using existing **R** packages.

Design of Experiments in R

Ulrike Grömping[1*]

1. Beuth University of Applied Sciences, Berlin

*Contact author: groemping@bht-berlin.de

Keywords: design of experiments
R packages: agricolae; AlgDesign; conf.design; DiceDesign; DiceKriging; DoE.base; DoE.wrapper; ICAOD; lhs; OptimalDesign; planor; rsm; tgp

This talk discusses the development of the landscape of packages on Design of Experiments in **R**, the current state of which is documented in the Experimental Design Task View (http://cran.r-project.org/web/views/ExperimentalDesign.html). That landscape is quite diverse, currently consisting of (at least) four larger areas and various specialized packages. Descendants and relatives of the pioneering packages *conf.desig*n (2001) and *AlgDesign* (2004) as well as the growing area of packages for computer experiments are considered in more detail, and recent additions based on modern optimization methods are discussed.

How to use R to hack the publicly available data about skills of 2M+ worldwide students?

Przemyslaw Biecek[1*]

1. University of Warsaw

*Contact author: przemyslaw.biecek@gmail.com

Keywords: visualisation; survey data; archivist; data mining
R packages: ggplot2; shiny; intsvy; archivist; knitr; BetaBit; PISA2012lite

During the talk I will introduce The Programme for International Student Assessment (PISA), an international survey that aims to evaluate education systems worldwide. It's a source of large data about educational performance and various other characteristics of over 2 000 000 students from 62 countries. The data from the last PISA assessment is available in the **R** package.

To play with it we will use packages and to explain the relation between parental occupation and student's performance. Then we will overview the package – the toolbox for statistical analyses of international

surveys. Finally we will discuss applications of packages and and the role of reproducibility and traceability of results. At the end I will introduce the project that aims to boost data science skills of students.

R and C++

Romain François[1*]

*Contact author: romain@r-enthusiasts.com

Keywords: C++; high performance; Rcpp; dplyr
R packages: Rcpp; dplyr

Rcpp is now considered a major success in the **R** community. It allows easy connection between **R** and C++. I'll review some of its history and use cases (e.g. *dplyr*).

Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm and its R interface

Marek Gągolewski[1*]

1. Systems Research Institute, Polish Academy of Sciences

*Contact author: marek@gagolewski.com

Keywords: hierarchical clustering; single linkage; inequity measures; Gini-index
R packages: genie

The time needed to apply a hierarchical clustering algorithm is most often dominated by the number of computations of a pairwise dissimilarity measure. Such a constraint, for larger data sets, puts at a disadvantage the use of all the classical linkage criteria but the single linkage one. However, it is known that the single linkage clustering algorithm is very sensitive to outliers, produces highly skewed dendrograms, and therefore usually does not reflect the true underlying data structure - unless the clusters are well-separated.

To overcome its limitations, we proposed a new hierarchical clustering linkage criterion called *genie*. Namely, our algorithm links two clusters in such a way that a chosen economic inequity measure (e.g., the Gini or Bonferroni index) of the cluster sizes does not increase drastically above a given threshold.

Benchmarks indicate a high practical usefulness of the introduced method: it most often outperforms the Ward or average linkage in terms of the clustering quality while retaining the single linkage speed. The algorithm is easily parallelizable and thus may be run on multiple threads to speed up its execution further on. Its memory overhead is small: there is no need to precompute the complete distance matrix to perform the computations in order to obtain a desired clustering. In this talk we will discuss its reference implementation, included in the *genie* package for **R**.

Simulation of complex synthetic data with the R package simPop

Matthias Templ[1⋆]

1. Vienna University of Technology

⋆Contact author: matthias.templ@gmail.com

Keywords: synthetic data; complex sample designs; population data; prediction
R packages: popSim

The production of synthetic datasets has been proposed as a statistical disclosure control solution to generate public use files out of protected data. This is also a tool to create "augmented datasets" to serve as input for micro-simulation models, and – more generally – the synthetic data sets can be used for design-based simulation studies in general. The performance and acceptability of such a tool relies heavily on the quality of the synthetic data, i.e. on the statistical similarity between the synthetic and the true population of interest. Multiple approaches and tools have been developed to generate synthetic data. These approaches can be categorized into three main groups: synthetic reconstruction, combinatorial optimization, and model-based generation. We introduce *simPop*, an open source data synthesizer. *simPop* is a user-friendly **R** package based on a modular object-oriented concept. It provides a highly optimized S4 class implementation of various methods, including calibration by iterative proportional fitting and simulated annealing, and modeling or data fusion by logistic regression, regression tree methods and many other methods.

Geo-located point data: measurement of agglomeration and concentration

Katarzyna Kopczewska[1⋆]

1. Faculty of Economic Sciences, Unievrsity of Warsaw

⋆Contact author: kkopczewska@wne.uw.edu.pl

Keywords: spatial location; agglomeration; specialization; SPAG; Ripley's K
R packages: rgeos; spdep; sp; rgdal; dbmss

Geo-located points, representing locations of business and other units can be analysed with regard to agglomeration and concentration patterns, which in fact indicate a density of region's coverage with the economic activity. There is plenty of measures on territorially aggregated data (so called cluster-based measures) and just few on geo-located individual data (so-called distance-based measures) as SPAG or Ripley's K. This is to present current possibilities of statistical analysis of geo-located data in **R**, as well its applications in regional science. It would consider the sensitivity of measures for different spatial patterns. Spatial package *rgeos* allows for treating points as geometries, what expands significantly the analytical capabilities. One can also see the neighbourhood relations between points, which can be applied further in econometric models.

Business 1

Turning R into production – make your models reality

Michal Zylinski[1⋆]

1. Microsoft

⋆Contact author: michalz@microsoft.com

Keywords: model deployment; Hadoop; scalability; interoperability; big data

Rapid popularity growth of **R** language among not only academic, but commercial communities poses also a few technical challenges. How to make **R** code interoperable and easily accessible to developer team? How to scale your models to sustain business needs? How to overcome initial hurdles, caused by introducing yet another technology stack? My talk will cover a few practical solutions that aim to resolve such difficulties in real-life scenarios.

Bringing R to Enterprise

Wit Jakuczun[1⋆]

1. WLOG Solutions

⋆Contact author: wit.jakuczun@wlogsolutions.com

Keywords: enterprise; big scale solution; productivity
R packages: knitr; checkpoint; h2o; doParallel; localsolver

Is **R** enterprise ready? How to manage development-test-deployment cycle? How to scale **R**? In this short talk I will present summary of our 10+ years experience in delivering big scale solutions with **R**.

Data science outside the box: Developing a generic scoring algorithm for customer acquisition

Erik Barzagar-Nazari[1⋆]

1. eoda GmbH

⋆Contact author: erik.barzagar-nazari@eoda.de

Keywords: data mining; customer acquisition; case study

One major task in virtually every predictive modelling project is to find the method best suited for the problem on hand. Luckily, most of the time data scientists can rely on one of the many already existing and well established methods such as Random Forests, Gradient Boosting Machines, Neural Networks or Support Vector Machines to solve a variety of regression and classification problems. However, in some cases these standard approaches are not directly applicable to the problem on hand and data scientists need to become creative. In this talk, we will present a case study about a project we recently conducted for the databyte GmbH in Germany. As one of the leading business information providers, databyte maintains a vast database of several million companies, containing information such as revenue, size and many other properties. The aim of the project was to develop an application which, after provided with the customer base of a databyte client, would be able to score companies in the databyte database in order to identify the most promising contacts for direct marketing campaigns. While developing this application, we were facing two major requirements: first of all, each client has its own customer base, hence we could not just train one model; in fact, the algorithm must be able to 'train itself' in every run. Secondly, the customer base only contains 'positive data', that means we are dealing with a so-called positive-unlabelled-problem.

Embedding R in business processes

Andreas Wygrabek[1⋆]

1. eoda GmbH

⋆Contact author: andreas.wygrabek@eoda.de

Keywords: business; process integration

There is no doubt about it: a variety of implemented statistical methods that are easy to access can make **R** a valuable tool for plenty application scenarios in business e.g. planning sales campaigns, monitoring machines, acceptance sampling on incoming goods or in-process control. All these scenarios are ruled by standardized business processes and tied with substantial financial risks. Beside the analytical capabilities of **R**, there are two main requirements on **R** that are important for companies: on the one hand **R** needs to be smoothly integrated into existing business processes. On the other hand, **R** needs to be a reliable part of the process chain: secure, easy to maintain, automatable and integrable into a right management system. Even though **R** reaches an incredible popularity many companies struggle to make **R** more than a proof of concept tool. The maturity level of **R** is reached when it has become crutial a part of business process chains. The talk will delight the main questions and critical issues of making **R** an essential link in a business process

chain. Furthermore the talk will present some state of the art strategies to face the depicted issues regarding security, reliability and automatization in operating systems.

Business 2

Enterprise R Platform – the what, the why and the how

Gergely Mark[1], **Csongor Somogyi**[1], **David Kun**[1⋆]

1. ownr.io; Functional Finances Ltd

⋆Contact author: david.kun@functionalfinances.com

Keywords: deployment; process; control; enterprise
R packages: base; roveR; shiny;

In this talk we will introduce a concept for an Enterprise **R** Platform, starting with expectations towards such a platform, followed by the benefits and finally giving a reference implementation architecture and a case study. **R** users in enterprise settings are often building shadow IT, with no controls but also no access to efficient code sharing, automation, deployment, version control, etc. This also means that others in the enterprise cannot benefit from the **R**-based solutions. Our proposed architecture addresses all of these issues by introducing a playground for the **R** users and a deployment process supporting all the flexibility needed while subject to regular controls. Our solution also includes a way to separate the **R** projects within a single environment, so different versions of the same package can be installed for instance. Finally, by providing a REST API, we enable non-**R** users to embed the **R**-based tools in any other application in the enterprise, ranging from MS Excel via Java-based ETL tools like Informatica and MI/BI tools like Business Objects up to and including in-house solutions created in other languages.

R in the Mittelstand: Bringing Data Science to small and mid-size companies

Oliver Bracht[1⋆]

1. eoda GmbH

⋆Contact author: oliver.bracht@eoda.de

Keywords: implementing R; R in business

Most success stories of **R** in business are told by either new economy companies or large international enterprises. However, there is another field where **R** can enable success. Small to midsize, mostly family

owned, old economy companies: the Mittelstand. Besides its size, Mittelstand companies are characterized by their organisational culture, their value sets and their management philosophies. Hence, the introduction of Data Science in general and **R** in particular to Mittelstand companies needs to consider this specific characteristic in order to be successful. This talk will focus on patterns that Mittelstand companies typically follow when they implement **R**. Main challenges of the process as well as road blockers on the way will be pointed out. It will give practical advice with real world examples and finally propose strategies of how to achieve long term success.

R tools and tricks for marketing inference in a big internet company

Paweł Ładyżyński[1⋆]

1. Grupa Wirtualna Polska

⋆Contact author: pawelladyz@wp.pl

Keywords: big data; data mining; predictive models; feature selection; deep learning; random forests; clustering; marketing inference
R packages: dplyr; shiny; data.table; ranger; Boruta; h2o;

To be send later.

Using Google Analytics with R

Michał Bryś[1⋆]

1. Allegro

⋆Contact author: michal.brys@gmail.com

Keywords: Google analytics; web analytics; Google Analytics API
R packages: plotly; ggplot2; RGoogle Analytics

How to download data from Google Analytics to **R** via API and create more than standard analysis using data visualisation and machine learning.

Using R for backtesting algorithmic trading strategies on high-frequency data

Piotr Wójcik[1⋆]

1. Faculty of Economic Sciences, University of Warsaw

⋆Contact author: pwojcik@wne.uw.edu.pl

Keywords: algorithmic trading; high-frequency data; backtesting; efficient calculations
R packages: xts; chron; quantmod; Rbbg; IBrokers; TFX; tseries; PerformanceAnalytics; caTools; TTR; inline; Rcpp; RcppArmadillo

Along with the advances in computer technology high-frequency trading developed in 1990s and became widely popular since then. Traders started to build algorithms that use highly developed quantitative models to automatically determine when and where to trade. The profitability of trading strategies based on such algorithms needs to be verified on historical data (backtested) prior to its application in real life. Intraday data in finance may have large volumes (up to 1440 minutes or 86400 seconds of quotations every day, i.e. 20 millions of observations for one year). In addition, developed algorithms are usually parametrized and their final version needs to be optimized with respect to these parameters. This introduces a possibly huge number of combinations that need to be compared with respect to selected performance measures. Large amounts of data and many variants of the algorithm require a computationally efficient tool that should also allow to relatively easily apply statistical models. And **R** together with C++ can provide such a tool. The presentation shows how **R** can be used to access intraday data and to develop and backtest different algorithmic trading strategies with the help of *Rcpp* family packages.

Data Workflow 1

Dynamic Inflation Rate Calculation of Fast-Moving Consumer Goods: Shiny-SparkR App

Olgun Aydin[1⋆]

1. Mimar Sinan University

⋆Contact author: olgunaydinn@gmail.com

Keywords: inflation rate; shiny; SparkR; shiny-SparkR; web scraping
R packages: SparkR; shiny; ggplot2; rvest

All official statistics centre or central banks of countries are calculating inflation rate in monthly period. Inflation is measured by the consumer price index inclueds the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services for specific intervals, such as yearly. The central banks and the statistics centres serve these information publicly. Most of the institutes serve the information monthly. Its hard to track Inflation rates for "Fast-Moving Consumer Goods(FMCG)" daily. To monitor this, I created *SparkR-shiny* App. The application is scraping top FMCG related web sites in daily based and stored the data on Spark stand alone cluster on Amazon Web Services (AWS) Elastic Compute Cloud (EC2), calculating consumer price index and day on day, month on month or year on year changes based on FMCG categories, visiualizing the results according to user defined criteria. With this application people could filter consumer price index for any time interval, any category also people could compare index changes of goods in different categories.

Introduction to RDruid

Michał Maj[1⋆]

1. Grupa Wirtualna Polska

⋆Contact author: Michal.Maj@grupawp.pl

Keywords: Druid
R packages: RDruid

This presentation will briefly describe *RDruid* package which is **R** connector for Druid data store.

Modules in R

Sebastian Warnholz[1*]

1. INWT-Statistics

*Contact author: wahani@gmail.com

Keywords: programming; functional-programming
R packages: modules; import; parallel

In this talk I present the idea of modules inside the **R** language. Modules are an organizational unit for source code. The key idea of this package is to provide a unit which is self contained, i.e. has it's own scope. The main and most reliable infrastructure for such organizational units of source code in the **R** ecosystem is a package. Compared to a package modules can be considered ad hoc, but – in the sense of an **R** package – self contained. Furthermore modules typically consist of one file; in contrast to a package which can wrap an arbitrary number of files. Inside of packages modules act more like objects, as in object-oriented-programming. In this talk I cover basic use cases in parallel computing and good coding practices.

Search phrases in e-commerce platform allegro.pl - big data analysis using SparkR

Liliana Pięta[1,2], **Mariusz Strzelecki**[1], **Anna Wróblewska**[1,3*]

1. Allegro
2. Cracow University of Economics
3. Warsaw University of Technology

*Contact author: anna.wroblewska@allegrogroup.com

Keywords: text mining; spark; big data
R packages: sparkR; wordcloud; tm; rms

R is a powerful tool for data visualization and analysis but it's speed is limited to only one CPU speed. It stays in the opposite with current Big Data trend that focuses on gathering and analysing huge, terabyte-scale datasets in clustered environment. Combining **R** and Spark technology we can synergize power of the both tools and we can efficiently work on huge datasets. During the presentation we will show you our approach to analyze search phrases in the biggest polish e-commerce marketplace platform Allegro.pl. We try text mining methods on phrases that our users type into the search box. Moreover we analyse transactions and user search events and try to find out the associations between searching phrase and successful transactional event. We will focus on challenges we solved during adapting *sparkR* in Allegro and share difficulties we experienced.

The full process of creating an R application in recommendation systems: from Dockerfile to Zabbix monitoring

Natalia Potocka[1*]

1. Grupa Wirtualna Polska

*Contact author: Natalia.Potocka@grupawp.pl

Keywords: R application; docker; monitoring
R packages: rkafka; elastic; httr; RJBDC; RZabbix

The presentation covers the main steps of creating a reproducible **R** application. First of all, there is a need of collecting the data. We present how to receive it from various sources, including Kafka, Elasticsearch or Hadoop (with the help of suitable **R** packages). After transforming the data, doing the necessary calculations and obtaining the desired outcome we show how to push out the results. When the application is ready and running, it needs to be controlled. We present the proccess of creating reliable monitoring (with the help of Zabbix and Hipchat notifications). All of that can be wrapped up in a Docker to make the application reproducible in different environments and platforms.

Data Workflow 2

Power of Java's RSession and rKafka in the data science team collaboration

Paweł Cejrowski[1⋆]

1. Grupa Wirtualna Polska

⋆Contact author: pcejrowski@gmail.com

Keywords: integration; microservices; rsession; kafka; team collaboration
R packages: rkafka

The times of single geniuses able to solve most advanced problems of the technical industry passed away. Nowadays, it is necessary to collaborate and support each other within teams. When it comes to Data Science teams, they are cross-functional and consist of people possessing different skills and having different duties. When working on a single product mathematicians, statisticians and engineers have to find right ways of exchanging and interfacing their piece of work. Unfortunately, not only they think differently, but they also use different platforms and languages. It implies in creating solutions for integration. One well known way of co-working is embedding code in different language in your own software. As an example, **R** code can be invoked from Java family languages using RSession. Another way of collaboration is passing messages throughout a middleware. Best known messaging system is Apache Kafka, originally developed by Linkedin. It is distributed among cluster nodes and allows high-throughput in a publish-subscribe manner and what is more important can be accessed from **R** code using *rKafka* package. In my presentation, I am going to familiarize you with those subjects and prepare for using them in production. There are a few tricks that will make your life easier.

R Shiny for Real-time Analytics and Insight Delivery – A Solution for Complex Data in Agriculture

Ming Shan[1⋆]

1. Kynetec

⋆Contact author: ming.shan@kynetec.com

Keywords: Bayesian; big data; choice modeling; complex survey; real-time; shiny; simulation; small area estimation; spatial; statistical learning; supervised; unsupervised; visualization
R packages: bayesm; cluster; dplyr; fields; ggplot2; kohonen; lme4; leaflet; magrittr; maps; maptools; party; rgdal; rgeos; sp; spatstat; shiny

One of the key challenges to successfully harness the value of big data is to uncover the critical insight through analytics and make it easily accessible to decision markers in an intuitive way. **R** *shiny* offers a platform for developing an end-to-end solution completely within **R** from mining the data with many modeling choices to insight delivery via powerful and dynamic data visualization. Such solution is highly suitable for agricultural data which is vast and dynamic, and increasingly abundant due to the recent development and adoption of precision farming and the related technologies. We would like to illustrate a *shiny* application that integrates a mixture of privately collected and governmental data across many key dimensions such as time, different geographic levels, crops and product types. Analytics including both supervised and unsupervised modeling capabilities and other approaches such as small area estimation, spatial analysis and complex survey estimation are included. Among many other specific examples is an application of discrete choice modeling using hierarchical Bayesian estimation to capture the preference of brand or product features and the price sensitivity by small geographic units and also allow the users to conduct market simulation by changing product mixture and prices. All the analytics are made and delivered in real-time.

Seamless external R server integration with Excel with step-by-step debugging of the R code

Adam Ryczkowski[1⋆]

1. statystyka.net

⋆Contact author: adam@statystyka.net

Keywords: Excelsi-R; Rstudio; Excel; VBA
R packages: Rserve; svSocket

Excel can be a great UI for repetitive, semi-interactive **R** procedures. I want to present (yet another) solution that allows using **R** commands in Excel. It features:

- ability to connect to external **R** server so you don't have to be limited by local RAM/CPU
- inspection of variables on external **R** server
- compatibility with RStudio with its awesome step-by-step debugging capabilities

This setup allowed me to create Excel front end for analysis of dataset with >4M cases and 400 variables.

Workflow around modelling in Data Science / R

Filip Stachura[1⋆], **Olga Mierzwa**[1], **Paweł Przytuła**[1]

1. Appsilon

⋆Contact author: filip@appsilon.pl

Keywords: workflow; process; data science
R packages: dplyr; caret; randomForest

Working as a data scientist usually means working with data, building models, evaluating the results, translating them into actionable insights, getting feedback from experts and repeating the process. Hardly ever one starts with the model that finally will be used in production. It is usually trial and error process of trying new things and experimenting. That's why data science project could get disorganized rapidly. Every test involves a new script, each script requires a multiple arguments and produces one or more data files. Keeping track of all this implied structure is a pain.

The talk stresses the importance of having a process around data science related tasks, while keeping the main focus on creating a data product. We demonstrate the implementation of the light data science workflow – Dataflows. Dataflows allows **R** users to create pipelines, without writing extra code, is self documenting and easy to start working with.

During the talk we show the benefits of formalizing and structuring the process of model building in **R**. We mention the bottlenecks, propose solutions and biggest wins accomplished by introducing Dataflows. We strive to share our hands-on experience from various data science projects using **R**.

Target audience practitioners, data scientists and researchers interested in rising standards of their current data product creation process, with the stress on reproducibility, early error detection, easy of results evaluation, comparison and communication with less technical colleagues.

Methodology 1

Predicting machine failures

Maren Eckhoff[1⋆]

1. McKinsey

⋆Contact author: Marta_Swiniarska@mckinsey.com

Keywords: failure modelling, machine learning

With the advent of the Internet of Things, companies collect a wealth of data that can be used to monitor degradation of their assets. In this talk, we will discuss how statistical modelling and machine learning can be used to predict failures and derive an optimal maintenance strategy. We will explore different modelling approaches and some of the feature engineering challenges. Moreover, the high value of combining different data sources to enrich the model will be explained.

Ensemble learning - idea and applications

Mateusz Filarowski[1⋆]

1. McKinsey

⋆Contact author: Marta_Swiniarska@mckinsey.com

Keywords: machine learning; ensemble learning; data mining
R packages: mlr

In every winning Kaggle solution nowadays you will find an ensemble technique. Latest and greatest methods of combining results from multiple learning algorithms can lead to better predictive performance. This lecture will give you an overview of techniques and real-world business cases.

Ensemble learning - implementation with mlr

Tomasz Smolarczyk[1⋆]

1. McKinsey

⋆Contact author: Marta_Swiniarska@mckinsey.com

Keywords: machine learning, ensemble learning, data mining
R packages: mlr

How to build multiple models and combine their results into powerful ensemble learners in less than one week? We will present a predictive modeling framework that is taking advantage of ensemble learning techniques and is heavily based on *mlr* package.

Methodology 2

Unattended SVM parameters fitting for monitoring nonlinear profiles

Emilio L. Cano[1⋆], **Javier M. Moguerza**[2], **Mariano Prieto Corcoba**[3]

1. The University of Castilla-La Mancha
2. Rey Juan Carlos University
3. ENUSA Industrias Avanzadas

⋆Contact author: emilio@lcano.com

Keywords: quality control; SVMs; nonlinear profiles
R packages: SixSigma; e1071; qcc

The monitoring of nonlinear profiles is a recent quality control technique. It allows to apply Statistical Process Control (SPC) methods to processes in which, rather than having a quality characteristic, there is a sort of nonlinear function that characterises the process. This method has been implemented in the *SixSigma* **R** package. The underlying idea is to compute a prototype profile and confidence bands using a data set from an in-control process, monitoring subsequent profiles thereafter. Thus, the same methodology used in well-known Shewhart control charts can be applied to complex processes. To this aim, raw data can be used. Nevertheless, using regularisation theory nonlinear profiles can be smoothed in order to better represent and analyse the profiles. In this work, we use Support Vector Machines (SVMs) to smooth profiles throughout the control process. Consequently, SVM parameters must be selected in order to reach a good fit of the nonlinear function at hand. Such parameters, namely: C and ϵ, can be explicitely assigned in the *smoothProfile()* function of the *SixSigma* package. However, a quality control practicioner seldom knows about SVMs, needless to say that they have no time to spend modelling functions. Hence, we rely on to automatically fit the SVM parameters using the process data, thereby achieving unattended SVM fitting. Furthermore, noise is previously estimated by means of a loess fit.

SimonsSVM: A Fast and Scalable Support Vector Machine Implementation for R

Philipp Thomann[1⋆], **Ingo Steinwart**[1]

1. ISA / University of Stuttgart

⋆Contact author: philipp.thomann@mathematik.uni-stuttgart.de

Keywords: support vector machine; machine learning; non-parametric classification; non-parametric regression
R packages: SimonsSVM

Support vector machines (SVMs) are non-parametric methods for various supervised learning scenarios like classification and regression. They have been studied extensively both theoretically and practically in the last 20 years. There are many well-known implementations also for **R**, for instance the package *e1071* provides bindings to *LIBSVM*. We present our recent package *Simons' SVM* for **R** with the following key features:

1. Unprecented speed:

- Compared to, e.g. *LIBSVM*, training that includes 5-fold cross validation on a 10x10 hyper-parameter grid is about 300 times faster on, e.g. a 1000 samples containing subset of the binary version of the classical covertype dataset.
- Partitioning strategies further decrease training (and testing) time without sacrifying generalization. For example, the full covertype data set (about 523.000 samples) takes less than 9 min.
- Partitioning even allows to attack huge problems. For instance the higgs data set (10 million samples) could be trained and tested in five hours.

2. Inclusion of some standard learning scenarios:

- (weighted) binary classification
- multiclass classification (both AvA and OvA)
- Neyman-Pearson-type classification
- Least squares / quantile / expectile regression

3. Flexible user interface:

- Fully integrated cross validation let's the user focus on parameters he/she can understand.
- The underlying implementation have more than twenty independent otions leading to an enormeous flexibility for the power user.
- Meaningful default values, which in many cases makes a fine-adjustment unneccesary.
- Comprehensive documentation.

Quick demo:

```
install.packages("SimonsSVM",
repos="http://www.isa.uni-stuttgart.de/software/R")
library(SimonsSVM)
d <- ttsplit(iris)
model <- svm(Species ~ ., d$train)
test(model, d$test)
```

More information:

`http://www.isa.uni-stuttgart.de/software/R/demo.html`

`http://www.isa.uni-stuttgart.de/software/R/documentation.html`

Discrete Choice Models in R

Daniel Guhl[1*], **Sebastian Gabel**[1]

1. Humboldt University, Berlin

*Contact author: daniel.guhl@hu-berlin.de

Keywords: discrete choice models; econometrics; Bayesian statistics; Stan
R packages: Stan; mlogit; gmnl; bayesm; ChoiceModelR; RSGHB

Discrete choice models (DCM) are a widely used class of models in economics, marketing, and transportation Science. These models are rooted in random utility theory and can be applied if a decision maker picks one alternative out of multiple options. The most commonly used variations of DCM are the multinomial logit model (MNL), the mixed logit model (MXL), and multinomial probit (MNP). Many **R** packages (e.g., *mlogit*, *gmnl*, *bayesm*) are available for frequentist and Bayesian inference of DCM. However, the different packages lack a unified data interface, common structure of functions and methods, and comparable output. Therefore, the analyst is faced with a typical dilemma: to test several models and follow established research processes (and maybe fulfill reviewer requests) one has to deal with data transformations, model translations, and output formatting, which is tedious, time consuming, and error-prone. In addition, the established **R** packages lack flexibility and transparency. To this end, we propose to use Stan, a general purpose modeling language for Bayesian inference written in C++ with interfaces to **R**, python, Matlab, Julia and Stata. We compare Stan-implementations of MNL and MXL models using simulated and real data. We show how Stan can be used for estimating more sophisticated models (e. g. models in willingness-to-pay space). We also provide a quick reference over our **R** package *DCM* that aims at closing the gap for existing packages regarding a unified data API and comparable model output.

Multivariate analysis of variance for functional data using R

Tomasz Górecki[1*], **Łukasz Smaga**[1]

1. Adam Mickiewicz University

*Contact author: tomasz.gorecki@amu.edu.pl

Keywords: functional data; multivariate analysis of variance
R packages: fda

We develop two testing procedures for multivariate analysis of variance problem for functional data. Similarly as the one-way analysis of variance for such data, this problem seems to be of practical interest. The first method approximates the functional data from each observational unit with of linear combination of orthonormal basis. Then time is integrated out from the usual MANOVA sum-of-squares and cross-products matrices. The null distribution of the standard MANOVA statistics are determined by permutation. In the second test, the functional data from each observational unit are projected on $\mathbb{R}^p$. Then, standard or permutation MANOVA tests are applied to the projected data. The main rationale for this test is that equality of the mean-functions vectors does not hold if equality of mean vectors does not hold for any random projection of the mean-function vectors. The performance of these methods is examined in comprehensive simulation studies. The results suggest that the tests can detect small differences between vectors of curves even with small sample sizes. We demonstrate how these methods can be performed efficiently in **R** by applying them to real world data. We implement these methods for **R** in our forthcoming package.

Text Mining in R

Christoph Hoffmann[1*]

1. Appstam Consulting GmbH

*Contact author: christoph.hoffmann@appstam.com

Keywords: text mining; classification; sentiment analysis
R packages: tm; e1071; twitteR

I would like to present the Text Mining cababilities of **R**. In particular the framework for Text Analysis with **R** provided by the *tm* package. Furthermore, I will show the publicly available API for Twitter. Then I will continue with a case study of Sentiment Analysis of Twitter Data/Feedback Data. This has been done with Naive Bayes Classification and/or Logistic Regression. The Case Study involves all the steps from obtaining raw text data to manipulating the data to running the actual algoritms adapted for text data. A bit of theory has been provided to motivate the use of the Bayesian classifier and if time allows I will run model comparisons.

Packages 1

Reconciling forecasts: the hts package

Rob Hyndman[1⋆]

1. Monash University

⋆Contact author: Rob.Hyndman@monash.edu

Keywords: forecast; time series
R packages: hts

Hierarchical time series occur when there are multiple time series that are hierarchically organized and can be aggregated at several different levels based on dimensions such as product, geography, or some other features. A common application occurs in manufacturing where forecasts of sales need to be made for a range of different products in different locations. The forecasts need to add up appropriately across the levels of the hierarchy. I will describe some new features in the *hts* package for **R** which provides several methods for analysing and forecasting hierarchical and grouped time series.

stplanr: an R package for transport planning

Robin Lovelace[1⋆]

1. University of Leeds

⋆Contact author: r.lovelace@leeds.ac.uk

Keywords: transport; GIS; modelling; visualisation
R packages: stplanr; leaflet; shiny

stplanr was developed to solve a real world problem: how to convert official data on travel patterns into geographic objects that could be plotted on a map and analysed using GIS? Over time the package has evolved to include a number of other functions. Analysis of road traffic casualty data, various routing algorithms, 'travel watershed' analysis and access to Google's Travel Matrix are all possible. This paper traces the development of these capabilities with a focus on applied case studies and reproducible examples.

R packages for social indicators

Łukasz Wawrowski[1*]

1. Poznań University of Economics and Business

*Contact author: lukasz8989@gmail.com

Keywords: sample surveys; social indicators
R packages: survey; laeken; ineq; vardpoor; convey

Social cohesion is a very popular motto in European Union. It is measured by many different indicators such as poverty rate, low work intensity indicator or quintile share ratio. They were established by the European Union as part of the Lisbon Strategy and the Europa 2020 Strategy. Its function is to control realization of above mentioned strategies. Presentation aims to compare **R** packages used in social indicators estimation. Features and efficiency of different packages will be examined.

brms: An R Package for Bayesian Multilevel Models using Stan

Paul-Christian Buerkner[1*]

1. The University of Münster

*Contact author: paul.buerkner@gmail.com

Keywords: Bayesian inference; multilevel model; MCMC, Stan
R packages: brms

The *brm* package implements Bayesian multilevel models in **R** using the probabilistic programming language Stan. A wide range of distributions and link functions are supported, allowing to fit – among others – linear, robust linear, binomial, Poisson, survival, ordinal, zero-inflated, hurdle, and even non-linear models all in a multilevel context. Further modeling options include autocorrelation of the response variable, user defined covariance structures, censored data, as well as meta-analytic standard errors. Prior specifications are flexible and explicitly encourage users to apply prior distributions that actually reflect their beliefs. In addition, model fit can easily be assessed and compared with the Watanabe-Akaike Information Criterion and leave-one-out cross-validation.

influence.SEM 2.0: An R Package for Sensitivity Analysis in Structural Equation Models

Gianmarco Altoè[1], **Massimo Nucci**[1], **Massimiliano Pastore**[1*]

1. University of Padova

*Contact author: massimiliano.pastore@unipd.it

Keywords: sensitivity analysis; influential cases; structural equation models
R packages: lavaan; influence.SEM

The **R** package *influence.SEM* 2.0 provides tools to perform sensitivity analysis in Structural Equation Models (SEM). Despite SEM are widely used by researchers in the social and behavioral sciences, the application of associated case-diagnostic tools (i.e. sensitivity analysis) have received limited attention and understanding in practice. Sensitivity analysis provides information regarding the impact of single cases on parameter estimates and goodness of fit of the model, thus supporting the researcher in the interpretation of results. In this paper, we present an easy-to-use **R** package yielding several measures of case influence for SEM via two applications to real data. We discuss the utility of detecting influential cases in SEM, provide recommendations for the use of measures of case influence, and give suggestions on how this analysis can be usefully employed in combination with other statistical techniques.

Packages 2

EnvCpt: An R package for changepoint identification in environmental data

Rebecca Killick[1*], **Claudie Beaulieu**[2], **Simon Taylor**[1]

1. Lancaster University
2. University of Southampton

*Contact author: r.killick@lancs.ac.uk

Keywords: model selection; changepoints; nonstationary time series; environment; oceanography; climate science
R packages: EnvCpt; changepoint

Man-made pressure on the Earths climate and ecosystems is increasing vulnerability to abrupt changes, which can be especially socio-economically challenging given the rapidity at which an ecosystem switches from one state to another relative to the time spent in the different states (e.g. a shift from one year to the next that persists on decadal or longer time scales). The ecosystem shift can be a response to change in external forcing (e.g. climate shift) or a random reorganization of the system, which can often be characterized by a simple autoregressive process. The distinction between stochastic and deterministic regime shifts is fundamental to gain a better understanding of the underlying mechanisms. In the climate and oceanography literature, the detection of regime shifts is commonly made using a sequential algorithm to test for shifts in the mean. However, this methodology can lead to spurious regime shift detection in the presence of autocorrelation, which is typically present in climate and environmental time series. Furthermore, a trend (e.g. long-term climate change) may be falsely interpreted as a series of regime shifts using this methodology. The *EnvCpt* **R** package presents a flexible methodology able to discern between the presence of mean, trend and autocorrelation and any combination of multiple shifts therein. The benefit of the package is that an automatic choice is made between whether trend, autocorrelation or changepoint models best fit the data using model selection. Thus no visual inspection of the data is required unlike current methods.

archivist 2.0: News from Managing Data Analysis Results Toolkit

Marcin Kosiński[1*]

1. Warsaw University of Technology

*Contact author: marcin.kosinski@grupawp.pl

Keywords: reproducible research; sharing objects; archiving
R packages: archivist; archivist.github

Open science needs not only reproducible research but also accessible final and partial results. During the speech I will present the most valuable applications of the *archivist* package. The archivist is an **R** package for data analysis results management, which helps in managing, sharing, storing, linking and searching for **R** objects. The *archivist* package automatically retrieves the object's meta-data and creates a rich structure that allows for easy management of calculated **R** objects. The *archivist* package extends the reproducible research paradigm by creating new ways to retrieve and validate previously calculated objects. These functionalities also result in a variety of opportunities such as: sharing **R** objects within reports/articles by adding hooks to **R** objects in table/figure captions; interactive exploration of object repositories; caching function calls; retrieving object's pedigree along with information on how the object was created; automated tracking of performance of models.

Reporting automation with ReporteRs

David Gohel[1⋆]

1. ArData

⋆Contact author: david.gohel@ardata.fr

Keywords: reporting; Word; PowerPoint; tables; graphics
R packages: ReporteRs; rtable; shiny

R usage has risen in companies and reporting capabilities of **R** is now an important subject. The way results are compiled in documents and spread to colleagues or customers can be time consuming. Key points are reproducibility, ability to produce pretty outputs and document types.

Microsoft document formats are still pretty ubiquitous in corporate environments. The package *ReporteRs* make easier the production of these documents on any platform. **R** users can produce pretty formatted outputs into Word or PowerPoint documents with only **R** code. It comes with an API to produce advanced graphical and tabular reporting.

In this talk, I will introduce *ReporteRs* and others related packages. Major features will be explained and illustrated. A use case will be presented to show a clinical reporting application made with *shiny*.

RevoScaleR - performance and scalability R

Łukasz Grala[1⋆]

1. Poznan University of Technology

⋆Contact author: lukasz@tidk.pl

Keywords: performance; scalability
R packages: RevoScaleR; RevoUtils; RevoUtilsMath; RevoMods; RevoTreeView; RevoPemaR

The Big Data scalability and performance are key issues. Microsoft Server **R** (HADOOP, Teradata, Spark) and **R** Services for SQL Server 2016 provides a set of libraries *RevoScaleR*. During the session, we compare the performance and scalability CRAN **R** and *RevoScaleR* in typical applications.

LimeRick: Bridge between LimeSurvey and R

Kamil Wais[1*]

1. University of Information Technology and Management in Rzeszów

*Contact author: kamil.wais@gmail.com

Keywords: LimeSurvey, on-line survey, CAWI, survey meta data
R packages: LimeRick (not publicly available yet)

The first public presentation of the prototype of a new **R** package that provides useful connection between **R** and the most popular open-source web scripts for on-line surveys (`http://www.LimeSurvey.org`).

The package aims to enable and simplify the work flow of reproducible CAWI research in Social Science; preparing templates for ad hoc and real-time analysis, performing detailed meta-analysis, archiving and monitoring responses directly from **R**.

With the *LimeRick* package one can import pre-processed data from an on-line survey with an **R** script (with the use of RemoteConrol2 API or with non-API solution). Then, the data can be processed analytically by an **R** script and automatically reported with the use of *shiny* package. The whole process can be pre-programmed and performed without the researcher interference. This enables to build data products based on declarative data from on-line surveys, which are processed, analyzed, and visualized on-line and in real-time. It can be of use for performing automatic tracking studies with real-time KPIs monitoring (e.g. for a customer satisfaction survey). The package will be also linked to a tool for unique, advanced analysis of meta-data from an on-line survey, which has already been alpha tested during large CAWI research project. As the early prototype of the package is presented, the presentation will end with a call for collaboration for developing and testing the package in different usage cases.

BioR

Are we ready for Personalized Medicine?

Paweł P. Łabaj[1⋆]

1. Department of Biotechnology, Boku University Vienna

⋆Contact author: pawel.labaj@boku.ac.at

Keywords: medicine; exposome; biomedical data, Bioconductor

In the era of fast-paced development of technology and services, there are limitless opportunities for customization to meet specific user needs. Over the next decade, as much as half of the proportion of health care will shift from the hospital and clinic to the home and community. With Personalized Medicine understood as prevention and treatment strategies that take individual variability into account we need to identify this individual variability via characterizing each person's individual baseline health state instead of resorting to population-based variable distributions. This health state baseline cannot be, however, determined with use of just the classical medical records. Recent technological advances have created opportunities to harness additional sources of biomedical data on a real time basis, for instance through the use of: (1) mobile medical devices for monitoring dedicated health parameters (insulin, heart rate, etc), and (2) wearables. The synergy of these two streams should provide a good estimate of the health state baseline. In order to model estimated data of health state baseline and future scenarios, it is imperative to include an important, yet largely missing third component - the exposome. This term cover all the exposures of an individual in a lifetime. So far it was mostly connected with air quality, light, climatic variations, ozone and volatile organic compounds. But we cannot forget about the 'living' component of exposome. As dense human environments such as cities account for over a half of the world population (in EU 80%) there is a need to build a molecular portrait of cities in order to study what lives around us and how it affects our health and wellbeing. There are, however, no dedicated tools, frameworks or standards how to store, share, integrate these streams of data. There is a lot to do as there are many open question and challenges which need to be address in order to provide valuable input. Fortunately, there is a very lively community around **R** Bioconductor project, which provides tools for the analysis and comprehension of high-throughput genomic and other biomedical data. This community should take a leading role in the future development of solutions for Personalized Medicine.

N-gram analysis of biological sequences in R

Michał Burdukiewicz[1⋆], **Piotr Sobczyk**[2], **Małgorzata Kotulska**[3], **Paweł Mackiewicz**[1]

1. Department of Genomics, University of Wrocław
2. Institute of Mathematics, Wrocław University of Technology
3. Department of Biomedical Engineering, Wrocław University of Technology

⋆Contact author: michalburdukiewicz@gmail.com

Keywords: n-gram; k-mer; feature selection; proteomics
R packages: biogram; ranger; slam

N-grams (k-mers) are vectors of n characters derived from input sequences, widely used in genomics, transcriptomics and proteomics. Despite the continuous interest in the sequence analysis, there are only a few tools tailored for comparative n-gram studies. Furthermore, the volume of n-gram data is usually very large, making its analysis in **R** especially challenging.

The CRAN package *biogram* facilitates incorporating n-gram data in the **R** workflows. Aside from the efficient extraction and storage of n-grams, the package offers also a feature selection method designed specifically for this type of data. QuiPT (Quick Permutation Test) uses several filtering criteria such as information gain (mutual information) to choose significant n-grams. To speed up the computation and allow precise estimation of small p-values, QuiPT uses analytically derived distributions instead of a large number of permutations. In addition to this, *biogram* contains tools designed for reducing the dimensionality of the amino acid alphabet, further scaling down the feature space.

To illustrate the usage of n-gram data in the analysis of biological sequences, we present two case studies performed solely in **R**. The first, prediction of amyloids, short proteins associated with the number of clinical disorders as Alzheimer's or Creutzfeldt-Jakob's diseases, employs random forests trained on n-grams. The second, detection of signal peptides orchestrating an extracellular transport of proteins, utilizes more complicated probabilistic framework (hidden semi-Markov model) but still uses n-gram data for training.

R as an Environment for the Reproducible Analysis of DNA Amplification Experiments

Stefan Rödiger[1⋆], **Michal Burdukiewic**[2], **Peter Schierac**[1]

1. Institute of Biotechnology, Brandenburg University of Technology Cottbus-Senftenberg
2. Department of Genomics, University of Wroclaw

⋆Contact author: stefan.roediger@b-tu.de

Keywords: bioinfrmatics; qPCR; digital PCR; reproducible research;non-linear regression; smoothing; pre-processingI;
R packages: chipPCR; dpcR; MBmca; RDML; rkwarddev; qpcR

Quantitative PCR (qPCR), digital PCR (dPCR), quantitative isothermal amplification (qIA) and melting curve analysis (MCA) are key technologies in molecular diagnostics, forensics and life sciences. However, software for the biostatistical data analysis of such technologies is either tied to a specific task or part of a

monolithic closed source software. This limits the reproducibility of computations and gives no control over the analysis algorithms.

We contributed bioinformatics software tools for reproducible and transparent data analysis. Analysis pipelines consisting of statistical procedures, raw data preprocessing, analysis, charts and report generation are implemented for increasingly demanded reproducible research. This is achieved with our packages *MBmca*, *chipPCR*, *dpcR* and *RDML*. We implemented selected findings for the **R** GUI/IDE RKWard and as *shiny* web browser applications. For rapid prototyping of RKWard plugins we used the *rkwarddev* package.

In our exemplary workflows we show analyzers for qPCR, qIA, MCA and dPCR experiments, which can be build in short development cycles. Our software is targeted at users who develop novel devices or users who wish to analyze raw and unprocessed data.

Big data genomics data warehouses analyses with R

Marek Wiewiórka[1⋆], **Tomasz Gambin**[1], **Michał Okoniewski**[2]

1. Institute of Computer Science, Warsaw University of Technology
2. ETH Zurich

⋆Contact author: marek.wiewiorka@gmail.com

Keywords: big data; genomics; data warehousing

Application of **R** in analyses of various kinds of genomic variants.

Using SparkR with distributed database in Parquet - a genomic example

Michał Okoniewski[1⋆]

1. ETH Zürich

⋆Contact author: michal.okoniewski@id.ethz.ch

Keywords: xxxxx
R packages: xxxxx

TBA

Methodology 3

Classic and network based cluster analysis: together we're better

Adolfo Alvarez[1⋆]

1. Analyx

⋆Contact author: adolfo.alvarez@analyx.com

Keywords: cluster analysis; networks; data visualization

This talk present the SAGRA method – a new cluster analysis methodology, which combines traditional and network based clustering techniques. Advantages of such combination are better performance and better visualization of results.

Is forest a pharmacy? - problems with data analyses

Marcin Dyderski[1,2⋆]

1. Institute of Dendrology, Polish Academy of Sciences
2. Department of Game Management and Forest Protection, Poznan University of Life Sciences

⋆Contact author: Marcin.Dyderski@gmail.com

Keywords: forest ecology; data mining; visualisation; linear models
R packages: caret; randomforest; ggplot2; hglm; sp; rgeos

Scientific supervisor: Dr. Andrzej M. Jagodzinski

Forest ecology is a specific branch of ecology, where due to extent spatial scale and large dimensions of individual objects we often need a special approaches. The aim of this presentation is to show a few case studies representing frequent analytical problems in forest ecology and its sources. I will present problem with unknown variability of studied phenomena on the example of individual biomass of herbaceous forest understory plant species. Although minimal sample size was unknown, the biggest analytical problem was discontinuity of species occurrences, connected with their specific biology. Thus, instead of two-factor analyses we performed simple ANOVA and exploratory data analysis to find specific patterns. In case of the study on alien species natural regeneration, due to high amount of zero values and lack of good explanatory

variables, we were forced to simplify our densities data into presence-absence and use logistic model to find interactions between type of tree stand and distance from the seed source. However, this interaction is hard to find without graphical data exploration. I also will present the study in coarser spatial scale, based mainly on published data on ancient woodland indicator species. This study is a simple analyses of species richness and use of random forest model to analyze highly-collinear variables – land-use types within grid square. Here I will show that simple (ANOVA) and more sophisticated (Poisson Hierarchic GLM with SAR) statistical approach lead into the same conclusions. I will also discuss potential sources of outliers and artifacts, which may be important threat for results interpretation.

k Prototypes Clustering of Mixed Type Data

Gero Szepannek[1⋆]

1. Stralsund University of Applied Sciences

⋆Contact author: gero.szepannek@fh-stralsund.de

Keywords: clustering; data mining; classification; k prototypes; multivariate statistics
R packages: clustMixType; klaR; cluster; fpc

Most literature in mutlivariate statistics deals with numeric data, same for clustering. In practical applications the analyst is often confronted with mixed type data of both numeric and factor variables. Hierarchical clustering can easily be extended to categorical data by specification of an apropriate (dis)similarity measure (e.g. Gower distance). The k means algorithm relies on euclidean distance. The *klaR* package offers an **R** implementation of the k modes extension to the k means algorithm for categorical data. Further, the recent *clustMixType* package also makes the k prototypes generalization to mixed type data acessible to **R** users. Both algorithms are presented together with their usage in **R**.

Education Learning

Revolutionize how you teach and blog: add interactivity

Filip Schouwenaars[1⋆]

1. DataCamp

⋆Contact author: filip@datacamp.com

Keywords: education; web technology; reporting; interactivity
R packages: tutorial; knitr; rmarkdown

R vignettes, blog posts and teaching materials are typically standard web pages generated with *rmarkdown*. DataCamp has developed a framework to make this static content interactive: **R** code chunks are converted into an **R**-session backed editor so readers can experiment. This talk will explain the inner workings of the technology, as well as a the tutorial **R** package that makes the transition to interactive web pages seamless. Some hands-on examples will showcase the remarkable ease with which you can convert **R** Markdown documents, vignettes and Jekyll-powered blogs into interactive **R** playgrounds.

Aargh I have to teach R (Experiences in the teaching of R)

Martin Schneider[1⋆]

1. eoda GmbH

⋆Contact author: martin.schneider@eoda.de

Keywords: teaching; training
R packages: shiny

As **R** is gaining popularity, so there is the interest in **R** courses. **R**, generally considered as a programming language with a steep learning curve, has surely its own typical pitfalls. The aim of this talk is to give an overview of the general do's and don'ts regarding the teaching of **R** and how a typical **R** course can be designed. Design and execution of a course depend on a variety of influences such as different types of course participants, content and context. Course participants usually differ regarding prior knowledge, interest in the material, ambitions as well as objectives to be reached. Generally, **R** courses focus either more on statistical or on programming topics. Another crucial aspect is whether the participants come from

an academic or an industrial background, in other words, the context in which the course takes place. The talk is based on the experiences of an **R** trainer teaching **R** to academics, as well as in companies, with a topic area ranging from teaching the fundamentals to specialised workshops (e.g. performing a Network Analysis in a *shiny* app).

Using R for artistic purposes

Päivi Julin[1⋆]

1. Aureolis

⋆Contact author: julin@apup.org

Keywords: visualization; design; typography; alternative; Processing
R packages: ggplot2; grid; Rserve; RColorBrewer

R can be tamed for serious business usage and data analysis, but this tool also has abilities for a different kind of approach: creating data art. Even **R** might not be mainly marketed with its graphics; it does have competence in genuine creativeness for unusual visualizations. Thus, **R** is much more than just bar plots, charts and maps – it is only a matter of practice and adaptation.

Data art focuses using raw data to produce visualizations with both aesthetic and informative scope. Having a creative process with digital media, real-time data and **R**, may lead to producing alternative interpretations from traditional information. Playing with **R** graphics devices, one is able to adjust any kind of visual presentations inside **R** console. Another option is to handle data manipulation with **R** and print output with alternative graphic portals - such as Processing or D3.

The presentation involves graphical visualizations, technical solutions, and general ideas, how to use **R** in alternative ways. Telling a story from data is a fun journey, therefore playing with **R** can be rather an amusing **R**-experience. This program being my muse, the aim is to inspire others.

Polish Diet Commissions - Text Analyzing

Kamil Krawczyk[1⋆], **Katarzyna Donaj**[1]

1. University of Adam Mickiewicz

⋆Contact author: krawczyk.grzegorz@gmail.com

Keywords: data mining; data visualization; text analyzing
R packages: ggplot2; shiny

We would like to talk about our investigation of texts from the general meetings of Polish Diet Commissions from two cadencies: the last one and the present one. We do believe that no one has ever done that before. Who knows what it can bring? We will discuss how we gathered the data and show some information that we get from it.

Lightning Talks

R as a tool for graphical diagnostics in population pharmacokinetic modeling

Agnieszka Borsuk[1⋆]

1. Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk

⋆Contact author: borsuk.agnieszka@gmail.com

Keywords: graphical model diagnostics; population pharmacokinetics
R packages: ggplot2; lattice

Population pharmacokinetic (PopPK) modeling aims at finding typical pharmacokinetic parameters and their variability within a target population of patients treated with a drug of interest. Population approach to pharmacokinetic modeling is gaining popularity as it can handle sparse data and estimate pharmacokinetic parameters of each individual. Large datasets and complexity of the models hinder assessing the quality of a model fit with a single numerical value. Graphical analysis plays a unique role in PopPK since it enables a better insight into the model structure. **R** is a powerful and versatile tool for graphical model diagnostics. It offers variety of visualizations, allows creating flexible scripts for quick model assessment and saving the results as a formal report. The produced graphics are publication-ready and meet the requirements for most non-standard and customized plots.

Machine learning modeling of phenological phases in Poland

Bartosz Czernecki[1⋆], **Jakub Nowosad**[2], **Katarzyna Jabłońska**[3]

1. Adam Mickiewicz University in Poznań, Department of Climatology
2. Adam Mickiewicz University, Department of Geoinformation
3. Institute of Meteorology and Water Management - National Research Institute

⋆Contact author: nwp@amu.edu.pl

Keywords: phenology; machine learning; modeling; satellite; meteorology; climatology
R packages: caret; raster; Boruta; parallel

Changes in timing of phenological phases are important proxies in contemporary climate research. The aim of the study was to create and evaluate different statistical models for reconstructing and predicting the

selected phenological phase within *caret* **R** package. Three types of data sources were applied as predictors:

i. satellite derived products,
ii. preprocessed gridded meteorological data
iii. spatial features (longitude, latitude, altitude) of the monitoring sites.

The obtained results has shown potential for coupling meteorological derived indices with remote sensing products in terms of phenological (late spring) modelling. It was also shown that choosing a specific set of predictors and applying a robust preprocessing procedures is more affecting final results than applying a statistical model.

Latent Class Analysis in Psychology

Paweł Kleka[1*]

1. Institute of Psychology, Adam Mickiewicz University

*Contact author: pawel.kleka@amu.edu.pl

Keywords: LCA; classification;
R packages: e1071

In my opinion, the methods of statistical analysis are not developing in psychology, it is happend just beyond. Then, methods are only used by psychologists on their home ground. Such a method, useful for building typology, is latent classes analysis (LCA), which I had 'discovered' by accident. In my presentation, I would like to show how use LCA to typical psychological data, which rarely meet the requirements of parametric methods (quantitative measuring scale, distribution normality, no outliers).

Exploratory data analysis of a clinical study group - revealing patient subgroups.

Bogumil M. Konopka[1*], **Felicja Lwow**[2], **Łukasz Łaczmański**[3]

1. Wroclaw University of Science and Technology
2. The University School of Physical Education in Wroclaw
3. Wroclaw Medical University

*Contact author: bogumil.konopka@pwr.edu.pl

Keywords: clustering; outlier detection; PCA
R packages: clv; ggplot2; chemometrics

Thorough knowledge of the structure of analyzed data allows to form precise research questions. The data structure and basic associations between parameters in the data can be revealed with methods for exploratory data analysis. Currently a researcher has a whole plethora of exploratory tools to choose from. Selecting methods that will work together well and facilitate data interpretation is not an easy task. In this work we present a well fitted set of tools for a complete exploratory analysis of a clinical study dataset and perform a case study analysis on a set of 515 patients. The proposed procedure comprises several steps:

1. robust data normalization,
2. outlier detection with Mahalanobis (MD) and robust Mahalanobis distances (rMD)
3. hierarchical clustering with Ward's algorithm,
4. Principal Component Analysis with biplot vectors.

Introductory analysis showed that the case-study dataset comprises two clusters separated along the axis of sex hormone attributes. Further analysis was carried out separately for male and female patients. The most optimal partitioning in the male set resulted in five subgroups. Two of them were related to diseased patients: diabetes and gonadotroph adenoma patients. Analysis of the female set suggested that it was more homogeneous than the male dataset. No evidence of pathological patient subgroups were found. In the study we showed that outlier detection with MD and rMD allows not only to identify outliers but can also assess the heterogeneity of a dataset. The case study proved that our procedure is well suited for identification and visualization of biologically meaningful patient subgroups.

Multidimensional Clustering of Web Analytics Data

Alexander Kruse[1⋆]

1. etracker GmbH

⋆Contact author: kruse@etracker.de

Keywords: data mining; clustering; web analytics
R packages: amap; dplyr; ggplot2

Web Analytics is the collection of data and their analysis regarding the behavior of website visitors. This presentation demonstrates a walkthrough on how to use multidimensional cluster analysis to divide a heterogeneous group of website visitors into smaller homogeneous segments. Focus is the appropriate selection of segmentation features, the determination of number of clusters and the usage of a multidimensional clustering technique. All preprocessing, analysis and visualization has been done with the **R** programming language.

Turning Text Mining into Language Mining: Corpus Linguistics in R

George Moroz[1⋆]

1. School of Linguistics, National Research University Higher School of Economics

⋆Contact author: agricolamz@gmail.com

Keywords: corpus linguistics; cross-linguistic linked databases; minority language analysis; text mining
R packages: shiny; rmarkdown; leaflet; stringr

Few books deal with both linguistics and R. Most of them are actually textbooks on statistics with examples of real-life linguistic problems solved. So it is about analyzing linguistic data in **R** (there is nothing special about it, statistically speaking), not about doing linguistics in **R**. Fortunately, there are some linguistic

packages in **R**. Santiago Barreda created package *phonTools* for phonetics research and experiments. Aaron Albin made a package *PraatR*, which provides **R** with the functionality of the well known phonetic program Praat by Paul Boersma and David Weenink. There are some text mining technics implemented in **R** (e.g. *tm*) and a few packages which allows extracting data from social networks.

However, linguistic **R** packages are mostly focused on phonetics. Pure text mining is insufficient for the linguistic research. So we decided to create an extended package, which helps to solve problems in linguistic typology and minority language description. In this talk, I will show preliminary results dealing with creation of a language corpora using *shiny*, *rmarkdown*, *leaflet*, and *stringr*. Problems, that I will cover during the talk:

- storing and analyzing texts in minority language: transcription, translation and morphological analysis; all examples are based on Adyghe (Circassian language family) and Mehweb (Northeast Caucasian language family)
- standardization of language documentation in **R**: integration with Cross-Linguistic Linked Databases

Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged

Structural bioinformatician's notebooks with pdbeeR and knitr

Paweł Piątkowski[1*]

1. International Institute of Molecular and Cell Biology

*Contact author: ppiatkowski@genesilico.pl

Keywords: bioinformatics; biomolecules; knitr; reproducible research
R packages: pdbeeR; knitr

Working with biological molecules often involves much work with .pdb files – analyzing, subsetting, transforming and visualizing structural data. Doing this by hand is tedious, hard to reproduce and prone to errors. A new **R** package – *pdbeeR* – can help you make these chores fun and save your work as elegant knitr notebooks.

Using R to incorporate data science into the undergraduate statistics curriculum

Paul Roback[1*]

1. St. Olaf College

*Contact author: roback@stolaf.edu

Keywords: undergraduate curriculum; statistical computing; data science
R packages: ggplot2; dplyr; ISLR

The American Statistical Association recently endorsed an updated set of Curriculum Guidelines for Undergraduate Programs in Statistical Science, and the first key point is: 'Increased importance of data science'. In addition to traditional topics in statistical reasoning, modeling, and inference, today's future statisticians also need the ability to access and manipulate data and to engage in algorithmic problem-solving to maximize their contributions. In this talk, I will explore what the increased importance of data science means for undergraduate programs in statistics and describe what we have done using **R** at St. Olaf College, an undergraduate institution of 3000 students in the United States. For example, statisticians partnered with colleagues in computer science to develop a 'statistics-infused' version of introductory computer science (CS125). This new statistical computing course introduces students to key foundational ideas in computer science by using mostly **R** (and some Python) and motivating the ideas through a data analysis context. Students learn about functions, loops, matrices, text strings, data scraping, and SQL while also discovering data visualization and classification methods. In fact, this course is one of seven featured in The American Statistician article 'Data science in the statistics curricula: preparing students to 'think with data". In addition to CS125, we have integrated more data wrangling and data visualization into our regression course, added a course on Algorithms for Decision Making which primarily explores classification methods using **R**, and developed an entire course on Data Visualization with **R** to pilot this fall.

Analysing the statistical effects of manipulated data

Luigi Lombardi[1*], **Marco Bressan**[1]

1. Department of Psychology and Cognitive Science, University of Trento

[*]Contact author: luigi.lombardi@unitn.it

Keywords: manipulated data; probabilistic models; generative models, sample generation by replacement

Many self-report measures collected in different research fields such as, for example, social and psychological studies, marketing and epidemiological studies may be affected by manipulated information by respondents. For example, an individual may deliberately attempt to manipulate or distort responses to simulate grossly exaggerated psychiatric symptoms in order to obtaining financial compensation or avoiding being charged with a crime. However, in other situations data manipulation can arise without a clear intention or goal by the respondent, for example, when an answer to a question is given under time pressure, or in heavily stressing contexts (e.g. during a job interview). Despite the reasons and motivations that induce people to (voluntarily or involuntarily) manipulate data can be different across contexts and situations, we believe that these processes can be represented by a single but flexible probabilistic model. In this contribution, we introduce an **R** package, called *sgr*, that can be used to perform data analysis on manipulated data according to a sample generation by replacement approach. The package includes functions for making simple inferences about discrete, ordinal as well as categorical data under one or more scenarios of data manipulation. The package also allows to quantify uncertainty in inferences based on possible manipulated data as well as to study the implications of manipulated data for empirical results.

Cryptography in R

Jeroen Ooms[1*]

1. UC Berkeley

*Contact author: jeroenooms@gmail.com

Keywords: encryption; signatures; openssl; elliptic curves
R packages: openssl; sodium

Introduce a few packages that implement cryptographic methods.

Visualizing changes in demographics with R

Piotr Sobczyk[1*]

1. Wrocław University of Technology

*Contact author: pj.sobczyk@gmail.com

Keywords: visualization, demographics
R packages: plotly; ggplot2; leaflet

How to show aging of society? What is a good way to visualize population projections? How to reveal inadequate public investments related to the process of suburbanization? Data visualization is a perfect tool to bring intuitions about those complex and massive problems closer to the people. In my lighting talk I explore several visualization challenges related to demographics. For each I propose a solution from my blog (http://www.szychtawdanych.pl). Examples I show are created with **R** packages *ggplot2*, *plotly* and *leaflet*. Codes are publicly available on my github http://www.github.com/psobczyk.

R for pharmacokineticists - smulation of steady-state concentrations of amiodarone in heart compartmental model as an example.

Zofia Tylutki[1*], **Sebastian Polak**[1]

1. Faculty of Pharmacy, Jagiellonian University

*Contact author: zofia.tylutki@doctoral.uj.edu.pl

Keywords: pharmacokinetics; modelling; parameter optimization
R packages: deSolve; FME

Amiodarone poses a known risk of torsade de pointes arrhythmia induction, according to CredibleMeds classification. Thus, the data on its effective concentration at heart tissue are desired. Patients prescribed to amiodarone usually are on long-term treatment following the standard dosing regimen, i.e. 200 mg t.i.d. as a priming dose, and 100 mg q.d. as a sustaining dose. The aim of the study was to build a functional heart model in order to simulate the amiodarone cardiac concentration at steady state. Three compartments of physiological volumes of plasma (central compartment), cardiac tissue, and pericardial fluid linked via first-order rate constants represented model structure. The model was described by set of differential equations,

and written in **R** v.3.1.3. Numerical solutions were computed using *deSolve* library. The model parameters were optimized using *FME* package. The fit was performed to mean amiodarone concentrations observed by Escoubet et al. in plasma and heart tissue, respectively, in 61 patients given amiodarone in a single dose. The optimized parameters ([h-1] ka: 0.038, kht_in: 0.200, kht_out: 0.149, kpf_in: 0.680, kpf_out: 9.700, ke: 5.000) were used to simulate amiodarone time-concentration profiles in two-months standard dosage schedule of amiodarone. An event function was describing multiple administrations. The simulated cardiac concentrations at steady-state were ca. 20 times higher than corresponding amiodarone levels in central compartment, which occurred to be in accordance with the ratios reported in literature that can be assumed to refer to steady state situation. Results support the feasibility of the model as well as the approach to simulate steady-state.

Poster Session

What are sampling errors in the vegetation studies using visual estimation of presence and cover of plants? R can help

Damian Chmura[1⋆], **Anna Salachna**[1], **Edyta Sierka**[2]

1. University of Bielsko-Biala
2. University of Silesia

⋆Contact author: dchmura@ath.bielsko.pl

Keywords: repeatability; interrater reliability, agreement; intraclass correlation coefficient
R packages: nlme; multilevel; irr; betapart; vegan

In vegetation science (e.g. phytosociology), visual estimates of plant cover belong to the most frequently used descriptors. The main reason for their attractiveness lies in the very low cost of the data obtained in this way, both in terms of labour, time and equipment. However, they are a subjective methods which result in sampling error, difficult to control. It is emphasised that using the visible estimation of cover requires some attempt and experience. However, many studies showed that comparison between observers yielded differences in the cover of recorded species as well as between repeated estimates of the same observer. We conducted several experiments with visual measurements of tree canopy, i.e. cover of tree layer and cover of herb layer in forest habitats. The experiments were performed with various raters, differing in experience in fieldwork: from students to professional researchers. Several field methods were applied: canopyscope, Braun-Blanquet approach, different scales of estimates; point method and a few vegetation types were chosen. Contrary to similar studies, where usually descriptive statistics are shown, we employed different statistical methods e.g. intra-class correlations, analyses of species pseudoturnover and nestedness – associated with loss of information among observes. The selected in **R** statistical methods turned out to be sensitive and robust tools. They revealed that even between professional researchers distinct differences appear. Advantages and disadvantages of available field methods are discussed and some possible improvements are suggested.

RNA-seq transcriptional profiling of PPD-b-stimulated peripheral blood from cattle infected with *Mycobacterium bovis*

Carolina N. Correia[1⋆], **Kirsten E. McLoughlin**[1], **Nicolas C. Nalpas**[1], **David A. Magee**[1], **John A. Browne**[1], **Kate E. Killick**[1], **H. Martin Vordermeier**[2], **Bernardo Villarreal-Ramos**[2], **Stephen V.**

Gordon[3], David E. MacHugh[1]

1. Animal Genomics Laboratory, UCD School of Agriculture and Food Science, University College Dublin
2. Animal and Plant Health Agency, UK
3. UCD School of Veterinary Medicine, University College Dublin

*Contact author: carolina.correia@ucdconnect.ie

Keywords: transcriptomics; differential gene expression; tuberculosis
R packages: AnnotationFuncs; magrittr; VennDiagram; MASS; RColorBrewer; svglite; org.Bt.eg.db; dplyr; edgeR; limma; extrafont; ggplot2; gridExtra

Mycobacterium bovis infection, the cause of bovine tuberculosis (BTB), costs an estimated $3 billion to global agriculture annually. During the last decade, the maturation of high-throughput sequencing technologies coupled with well-annotated genome resources, has provided an unprecedented opportunity to gain a deeper understanding of host-pathogen interactions for many infectious diseases. Within this context, transcriptional profiling of the host immune response to *M. bovis* infection is a powerful approach for identifying host genes and cellular pathways important to disease pathology. For the present study, ten age-matched male Holstein-Friesian calves were infected endobronchially with *M. bovis* (~2,000 CFU – colony forming units). Peripheral blood samples were collected in duplicate at four time points (-1 week pre-infection, +1 week, +2 week, and +10 week post-infection) and used for:

a. an overnight stimulation with purified protein derivative of bovine tuberculin (PPD-b) at 37°C
b. a control overnight incubation at 37°C without PPD-b stimulation.

After isolation of total RNA, poly(A)+ purified RNA was used to generate strand-specific RNA-seq libraries for high-throughput sequencing. Transcripts were quality checked, adapter and quality filtered, and then aligned to the *Bos taurus* reference genome UMD3.1.1. Following summarisation of gene counts, lowly expressed transcripts were removed prior to subsequent gene annotation and differential expression analyses. Results showed 929 differentially expressed (DE) genes at -1 week pre-infection, 1,619 DE genes +1 week post-infection, 1,170 DE genes at +2 week, and 5,535 DE genes at +10 week (compared to non-PPDb-stimulated group at each time point; FDR correction threshold ≤ 0.05).

Pharmacokinetics-driven modeling of metabolomics data

Emilia Daghir-Wojtkowiak[1*], Paweł Wiczling[1], Małgorzata Waszczuk-Jankowska[1], Roman Kaliszan[1], Michał J. Markuszewski[1]

1. Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk

*Contact author: emilia.daghir@gmail.com

Keywords: pharmacokinetics; metabolomics; Bayesian analysis; ROC; cancer
R packages: runjags; coda; caTools; MCMCglmm; caret; Hmisc; lattice

Metabolomics is a dynamically developing research area utilizing high-throughput analytical techniques to detect spectrum of changes in metabolites' levels. However, technological improvements designed for analytical systems are not parallel with the development/application of computational methods which would allow for more efficient elucidation of knowledge from large datasets. In this study, we introduced the concept of pharmacokinetics-driven modelling of targeted metabolomics data comprising nucleoside and creatinine concentration measurements in urine of healthy and cancer patients. An approach using Bayesian analysis

was used for the estimation of model parameters. The classification performance of the proposed model was summarized via area under the ROC (Receiver-Operator Curve), sensitivity and specificity using external validation. Cancer was associated with an increase in methylthioadenosine/creatinine excretion rate by a factor of 1.82 (1.33–2.47). Age influenced nucleosides/creatinine excretion rates for all nucleosides in the same direction with likely sex-related differences among several nucleosides' concentrations. The individual a posteriori prediction of patient classification as area under the ROC was 0.58 (0.5-0.68) with sensitivity and specificity of 0.63 (0.46–0.76) and 0.58 (0.45–0.68), respectively suggesting limited usefulness of 13 nucleosides/creatinine measurements in predicting the disease in this population. Pharmacokinetic-based approach in metabolomics may be useful in understanding the data when searching for potential disease indicators.

R as a tool for geospatial modeling in large dataset - example of dasymetric modeling at a continental scale (United States)

Anna Dmowska[1⋆]

1. Adam Mickiewicz University

⋆Contact author: dmowska@amu.edu.pl

Keywords: geospatial data; large dataset; dasymetric modeling
R packages: rgrass7; rgdal; RSQLite; sp

Geospatial raster datasets are usually large data structures. Thus its processing requires algorithms that are both efficient and fully automatic. The philosophy of development of GIS software offers many solutions to a user. These solutions are conceptually limited which means that they perform well in their native tasks but are difficult to extend if the problem goes beyond its foundations. In such situations computationally complex tasks require tailored software which is difficult to develop without programming experiences. **R** language offers extensive amount of tools designed to work with geospatial data like *sp* library, as well as bindings to external data sources like *rgrass7*, *rgdal*, *RSQLite*. Flexible nature of geospatial objects in **R** allows to create a fully automatic procedure that complements GIS software without long-term programing experiences.

Here we present an automatic procedure which was designed to work on high resolution datasets at a continental scale: 11 million of records in tabular data and over 8 billion of grid cells. To process such amount of data we automated our procedure by dividing it into few steps:

1. import and store Census data in SQLite database,
2. preprocessing of geospatial data performed in GRASS GIS software,
3. performing dasymetric modeling in **R**,
4. propagation model for geospatial data and export geospatial data from **R** to GIS format to map population distribution.

Our results show that:

1. **R** can perform such a calculation in a reasonable time (60 hours for the entire U.S.),
2. no additional intermediate layers or steps are required,
3. procedure is flexible: it can be reproduced regardless of the data scale.

Application of Artificial Neural Network to Planar Chromatography Data

Dimitri Fichou[1*], **Gertrud Morlock**[1]

1. Justus Liebig University Giessen, Germany

*Contact author: dimitrifichou@gmail.com

Keywords: chemometrics; neural network; planar chromatography
R packages: deepnet; jpeg; abind; EBImage

Planar chromatography has a unique specificity compared to other chromatographic techniques, i. e. the data format is a picture, in which each pixel has quantitative and qualitative properties that corresponds to the molecular reality in the physical word. This results in a greater amount of data points that allows the use of high-level machine learning algorithms like artificial neural network. Restricted Boltzmann Machine was used on planar chromatograms for denoising and classification. For both, no other preprocessing than a normalization between 0 and 1 was needed. The denoising task took as input patches of pixels; when crossing the network, the noise is removed and only the signal remains. For the classification task, several layers of this neural network were stacked together in order to analyze vertical bands of pixels. The last layer of this network discriminated the two classes of the dataset with an accuracy of 85 % when compared to human decisions.

cgmisc: enhanced genome-wide association analyses and visualization

Jagoda Jabłońska[1*], **Marcin Kierczak**[2], **Simon Forsberg**[3]

1. University of Warsaw
2. Uppsala University
3. Swedish University of Agricultural Sciences

*Contact author: jagoda100jablonska@gmail.com

Keywords: GWAS; bioinformatics; visualization; sequencing
R packages: cgmisc; GenABEL

High-throughput genotyping and sequencing technologies facilitate studies of complex genetic traits and provide new research opportunities. The increasing popularity of genome-wide association studies (GWAS) leads to the discovery of new associated loci and a better understanding of the genetic architecture underlying not only diseases, but also other monogenic and complex phenotypes. Several softwares are available for performing GWAS analyses, **R** environment being one of them. We present *cgmisc*, **R** package that enables enhanced data analysis and visualization of results from GWAS. The package contains several utilities and modules that complement and enhance the functionality of the existing software. It also provides several tools for advanced visualization of genomic data and utilizes the power of the **R** language to aid in preparation of publication-quality figures. Some of the package functions are specific for the domestic dog (*Canis familiaris*) data but are easy to adjust for analysing other species.

Penalized regression inference regarding variable selection in regular and high dimensions: comparison of selected methods implemented in R

Marta Karas[1⋆]

1. Department of Biostatistics, Indiana University Fairbanks School of Public Health

⋆Contact author: marta.karass@gmail.com

Keywords: penalized regression; inference; p-value
R packages: lassoscore; hdi; refund

In recent years, penalized regression techniques, such as lasso, have become commonly used for variable selection in linear regression modeling, but relatively little work has been done on quantifying the uncertainty in these procedures. Here, we perform comparison of four different approaches to this problem. With a score test method, penalized regression of an outcome on all but a single feature is performed, and test for correlation of the residuals with the held-out feature follows. PEER Ridge estimation procedure exploits the equivalence between penalized least squares estimation and a linear mixed model representation, and thus provides an automatic selection of tuning parameter alpha using REML criterion; then, the General SVD provides algebraic insight and a convenient way to derive the variance expressions of the estimates, which allows to perform inference regarding variable selection. With Multi sample-splitting method, the sample is split into two equal halves, where the first half is used for variable selection and the second half, with the reduced set of already pre-selected variables, is used for statistical inference in terms of p-values. In stability selection method, a stable subset of variables is selected in stochastic simulation where probability of a variable being selected into a model is approximated. We utilized **R** implementations of the methods listed above to compare variables selection inference in terms of FDR and power. We experimented with different covariance settings as well as dimensions of X design matrix in looking for situations when one method might be preferred.

Wrestling with big data in forestry: use of R in Scots pine site index analysis.

Wojciech Kędziora[1⋆], **Robert Tomusiak**[2]

1. Department of Forest Management Planning, Geomatics and Forest Economics, Warsaw University of Life Sciences – SGGW
2. Laboratory of Dendrometry and Forest Productivity, Warsaw University of Life Sciences – SGGW

⋆Contact author: wojciech.kedziora@wl.sggw.pl

Keywords: national forest inventory; permanent plots; site index

National Forest Inventory (NFI) supplies high resolution data on forests of all property forms located throughout the country. These data are collected from the circular sample plots uniformly distributed in grid of squares 4 by 4 km. In Poland, around ~27,900 sample plots were established. Recently, the second measuring cycle NFI was completed, which, in addition to a potentially wide range of spatial analysis, creates a unique opportunity to analyze the change in the pattern of tree growth through time, giving opportunity to describe the reaction of species to changes of environmental conditions. NFI data give a great

opportunity to determine one of the most important forest habitat's productivity factors – site index. It is often expressed as the average height of the trees of the target species at a given age. Determination of the site index for the dominant species in the stand allows one to characterize its growth potential in examined habitat conditions. The project envisages the use of empirical data collected in NFI to determine the structure of pine stands site index in Poland. Relatively high number of sample plots and complicated computations as well as future need for geostatistical analyses were reasons to choose **R** environment to work successfully with forestry big data.

An R implementation of Kauffman's NK model

Tomasz Owczarek[1⋆]

1. Faculty of Organization and Management, Silesian University of Technology

⋆Contact author: tomasz.owczarek@polsl.pl

Keywords: nk model; complexity; fitness landscape; agent-based modeling
R packages: dplyr; ggplot2; igraph

Stuart Kauffman's NK model is a simple but powerful model of complex system with many interactions between its elements. It generates so called fitness landscape, i.e. a space consisted of many points with different attractiveness (fitness). The great advantage of the model is that only two parameters (N and K) are required to be controlled to generate environments with 'tunable complexity' – from very simple with one local and global optimum to completely random, with a large number of local optima. The NK model is used in the studies of many phenomena in various fields e.g. biology, economics, organizational studies or engineering. In my presentation I would like to introduce an **R** package which implements the NK model and show how to use it to generate fitness landscapes and explore their properties for different influence matrices.

R as an effective data mining tool in chemistry

Anna Rybińska[1⋆], **Katarzyna Odziomek**[1], **Tomasz Puzyn**[1]

1. Laboratory of Environmental Chemometrics, University of Gdansk

⋆Contact author: rybinska@qsar.eu.org

Keywords: data mining; predictive modeling; ionic liquids; chemometrics
R packages: ggplot2; stats; dendextend; matrixStats; Matrix; MASS; klaR; cluster; dplyr; FWDselect

The vast amount of digital information generated every day in social media, industry and academia necessitates the use of advanced techniques enabling the processing, analysis and interpretation of data. A significant amount of data is generated during various types of chemical experiments. In this work, we present a workflow for predictive modeling of physical-chemical properties of a diverse group of chemicals (ionic liquids). Using selected chemometric methods, available in the most popular **R** packages, we analyze the chemical data, model key physical-chemical properties and validate the results. Moreover, the presented

approach enables the evaluation of the raw data quality. We choose two unsupervised methods for data exploration: hierarchical cluster analysis (HCA) and principal component analysis (PCA). Multiple linear regression technique (MLR) was chosen as the example of a modeling technique.

Applying genetic algorithms to calibrate a processing chain for a Landsat-based time series analysis of disturbance - regrowth dynamics in tropical forests

Fabián Santos[1*], **Gunther Menz**[1], **Olena Dubvyky**[1]

1. University of Bonn

*Contact author: fabian_santos_@hotmail.com

Keywords: time-series analysis; genetic algorithms; Landsat; tropical forests monitoring
R packages: GA; changepoint; BreakoutDetection; ecp

We present an innovative approach to calibrate processing chains designed for a time-series analysis based on genetic algorithms (GA). Using a case study of disturbance-regrowth monitoring employing Landsat data in tropical forests located in the Amazonian region of Ecuador. This area is characterized by mountainous terrain, high topographic relief and extensive cloud cover. These conditions can cause noise in the time-series of Landsat data and require different corrections before its analysis. We describe here the use of *GA* to reduce calibration uncertainties, enhance the output accuracy, and avoid unnecessary processing caused by trial-and-error approaches in remote sensing.

Modelling the distrubution of the bryophytes in different spatial scales

Sylwia Wierzcholska[1*], **Marcin K. Dyderski**[2,3], **Andrzej M. Jagodziński**[2,3]

1. Białowieża Geobotanical Station, University of Warsaw
2. Institute of Dendrology, Polish Academy of Sciences
3. Department of Game Management and Forest Protection, Poznań University of Life Sciences

*Contact author: sylwia.wierzcholska@gmail.com

Keywords: species distribution model, biodiversity, plant ecology, predictive modelling, mosses, forest
R packages: dismo, rasterVis

Species distribution models (SDM) are mathematical models describing distribution of species within environmental (ecological) and geographical space. These models are used for biodiversity conservation in two ways: to find out threatened species requirements and to predict spread of invasive species. We aimed to model distribution of model bryophyte species – *Dicranum viride* – in Poland and in Europe and compare its ecological niches obtained by these two models. We chose *D. viride*, as this easily-recognizable species is a subject of Natura 2000 protection and an ancient forest species. Thus, it may be an umbrella species for numerous bryophyte species occurring on decaying wood and in old woodlands. We used data

from Global Biodiversity Information Facility (`http://www.gbif.org`), published papers and herbarium collections to find out complete information about *D. viride* localities. As most of data about species distribution for large areas is presence-only data, we used MaxEnt model from *dismo* package, which is developed to processing this type of input data. As the explanatory variables we used 19 bioclimatic statistics from WordlClim database, available in 2.5' grid and, in case of model for Poland only, data about share of old (>100 years old) forests within grid square. We also analyzed data about phorophyte species and collection year. Model developed by MaxEnt is a probalistic model, which due to Receiver-Operator Curve allow to manage the threshold of species occurence, and thus – its restrictiveness. Our model has shown importance of particular bioclimatic variables and potential distribution of species. Obtained results allow us to conclude about climatic requirements of species studied and its potential habitats, where species may be found or may be protected ex situ. SDMs are very usefull tool for plant geography, biodiversity conservation and ecology.