R Training

Education of R - from the mere basics to the necessary skills to use and deploy R and advanced analytics in your organisation

www.bnosac.be – info@bnosac.be

bnosac

open analytical helpers

Overview

This document gives an overview of the training offered by BNOSAC regarding the use of **R (www.r-project.org)**.

The training consists of a set of modules which can be chosen depending on your skillset and ranges from introductory to advanced R programming, package development, R deployments and R analytics (statistical learning, text mining, plotting, graphics, spatial statistics) as well as on-demand training. Courses can be given throughout the year, on request. If interested, do not hesitate to get into contact.

> ⓘ Some of the courses are offered also through **LStat** – the Leuven Statistics Research Center on a yearly basis, other courses are offered at the **BNOSAC** offices or can be given at **your offices**. The courses about the use of Oracle R Enterprise are offered in cooperation with our **Oracle Partner Tripwire Solutions**.

Courses

The following courses are offered by BNOSAC

- **R programming**
  - R for starters
  - Common data manipulation for R programmers
  - Reporting with R
  - Creating R packages and R repositories
  - Managing R processes
  - R code management, Git and Continuous Integration
  - Integration of R into web applications
- **R analytics**
  - Statistical Machine Learning with R
  - Text mining with R
  - Applied Spatial modelling with R
  - Big data analytics with R on top of Hadoop, Spark & HAWQ
  - Computer Vision with R and Python
- **Oracle R Enterprise**
  - ROracle and Oracle R Enterprise (ORE) - transparancy layer
  - Oracle R Enterprise - advanced data manipulation
  - Data mining models inside Oracle R Enterprise (ORE) and Oracle Data Mining (ODM)

Price

The prices of the courses are indicated as a price per person per day. All courses except the courses on the use of Oracle R Enterprise are offered at the same price per day.

- The price of **Oracle R Enterprise courses is set to 375€ (excl. VAT) per day per person**.
- The price of all the other courses is **300€ (excl. VAT) per day per person – a half a day consists of 150€ (excl. VAT) per person.**

Contact

For more information about specific course content or to request the course in your offices, you can get into contact with BNOSAC.

Jan Wijffels
jwijffels@bnosac.be
+32 486 611708
www.bnosac.be

Courses on R as a programming language

Programming module I: R for starters

Course duration: **2 days**

What is covered in this module:

- What is R, packages available (CRAN, R-Forge, ...), R documentation search, finding help, RStudio editor, syntax
- Data types (numeric/character/factor/logicals/NA/Dates/Times)
- Data structures (vector/data.frame/matrix/lists and standard operations on these)
- Saving (RData) & importing data from flat files, csv, Excel, Oracle, MS SQL Server, SAS, SPSS
- Creating functions, data manipulation (subsetting, adding variables, ifelse, control flow, recoding, rbind, cbind) and aggregating and reshaping
- Plotting in R using base and lattice functionality (dot plots, barcharts, graphical parameters, legends, devices)
- Basic statistics in R (mean, variance, crosstabs, quantile, correlation, distributions, densities, histograms, boxplot, t-tests, wilcoxon test, non-parametric tests)

During the course, you will need to do exercises, so bring your laptop.

Introduction to R

b nosac

open analytical helpers

Jan Wijffels:
jwijffels@bnosac.be

BNOSAC - Belgium Network of Open Source Analytical Consultants
www.bnosac.be

Programming module II: Common data manipulation for R programmers

Course duration: **1 day**

This module allows you to be a better programmer by writing your own functions, getting acquainted with commonly used R functions for basic data manipulation and the R object oriented programming environment.

The following is covered in this module:

- with, within, by, apply family of functions & split-apply-combine strategy
- vectorisation, parallel execution of code
- data.table - fast group by, joining and data.table programming tricks
- basic regular expressions
- writing your own functions
- do.call
- reshaping from wide to long format
- environments
- S3 classes, generics and basic S4 methodology
- handling of errors and exceptions, debugging code

Be prepared on some tough exercises, so bring your laptop. A part of this course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Programming module III: Reporting with R

Course duration: **1 day**

If you want to create a report using R, either static, dynamic or an interactive web page, R provides several tools to do this. This module teaches you the basics of building reports with R. It covers the following topics.

- Sweave & knitr
- Markdown & pandoc
- integration with MS Office & presentations
- making R package vignettes
- An introduction to Shiny and interactive HTML reporting

A part of this course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Programming module IV: Creating R packages and R repositories

Course duration: **1 day**

R is notorious for it's flexibility by having more than 6000 packages available for direct usage. If you want to create your own package to distribute code to others inside your organization, this module teaches you how to build your own package and set up an enterprise R package repository.

- structure of an R package
- documenting your code and your R package using roxygen in RStudio
- check, build, install your R package
- unit-testing your R code
- creating your own R package repository

A part of this course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Programming module V: Managing R processes

Course duration: **1 day**

R is a programming language and can be launched from the command line. This module will learn you how to launch R properly in order to automate processes

- R and Rscript and the options
- Startup scripts and settings
- Handling of command line arguments
- Understanding package library folders
- Automating R processes in Windows & Linux
- Handling and logging of error/warning messages

Programming module VI: R code management, Git and Continuous Integration

Course duration: **1 day**

If you are an R developer and want to make sure your analysis is reproducible and traceable, you need to learn how to use a code repository. In this session, you will learn how to use RStudio for code maintenance. The following items are covered:

- R Coding guidelines
- Good practices in management & structuring of R code
- Git & code repositories
  - Setup
  - Push, pulling and cloning code
  - Handling of conflicts
  - Setting branches and releases
  - Use of RStudio's IDE to integrate with your code repository
- R package repositories
  - Build a local CRAN repository with your own packages
- Doing continuous integration with Gitlab
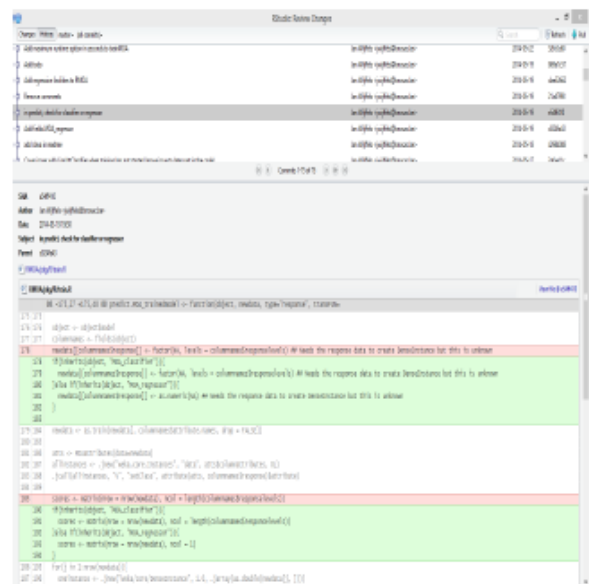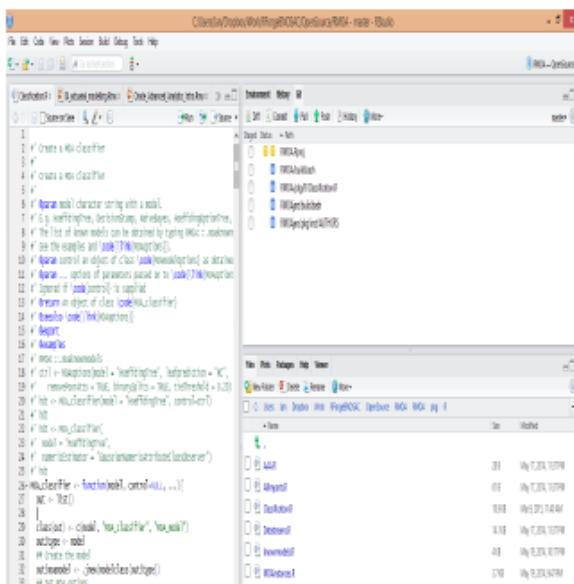
Programming module VII: Integration of R into web applications

Course duration: **2 days**

Learn how to create basic web applications and web services in R. The following elements are covered in the course:

- Setting up Shiny applications
- RApache & OpenCPU
- using R alongside javascript, htmlwidgets
- setting up webservices using R
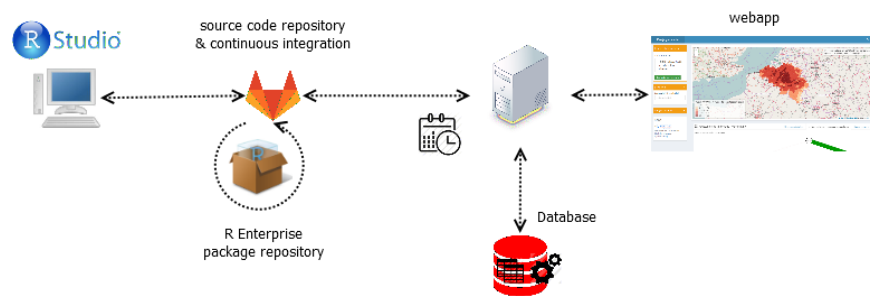
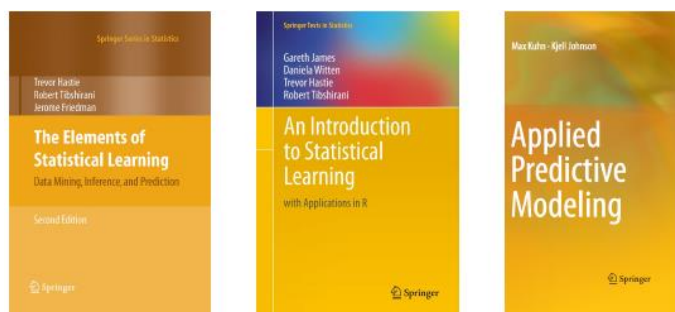# Analytical courses on R usage & applications

Analytical module I: Statistical Machine Learning with R

Course duration: **2 days**

This course is a hands-on course covering the use of statistical machine learning methods available in R. The following basic learning methods will be covered and used on common datasets.

- classification trees (rpart)
- feed-forward neural networks and multinomial regression
- random forests
- boosting for classification and regression
- bagging for classification and regression
- penalized regression modelling (lasso/ridge regularized generalized linear models)
- model based recursive partitioning (trees with statistical models at the nodes)
- training and evaluation will be done through the use of the caret and ROCR packages

This course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

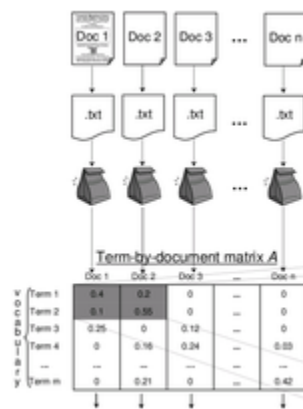## Analytical module II: Text mining with R

Course duration: **2 days**

This course is a hands-on course covering the use of text mining tools for the purpose of data analysis. It covers basic text handling, natural language engineering and statistical modelling on top of textual data. The following items are covered.

- Text encodings
- Cleaning of text data, regular expressions
- String distances

- Graphical displays of text data
- Natural language processing: stemming, parts-of-speech tagging, tokenization, lemmatisation
- Sentiment analysis
- Statistical topic detection modelling and visualization (latent diriclet allocation)
- Automatic classification using predictive modelling based on text data
- Visualisation of correlations & topics
- Word embeddings
- Document similarities & Text alignment

This course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Analytical module III: Applied Spatial modelling with R

Course duration: **2 days**

This course is useful for data scientists and data analysts which work frequently with data with a spatial component (data with latitude/longitude information). It gives an introduction to the numerous spatial facilities of R and some standard spatial statistical models. The following items are covered during the course.

- The sp package to handle spatial data (spatial points, lines, polygons, spatial data frames)
- Importing spatial data and setting the spatial projection
- Plotting spatial data on static and interactive maps
- Adding graphical components to spatial maps
- Manipulation of geospatial data, geocoding, distances, …
- Density estimation, kriging and spatial point pattern analysis
- Spatial regression

This course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Analytical module IV: Big data analytics with R on top of Spark & Hadoop

Course duration: **2 days**

This course is useful for data scientists and data analysts which work frequently on big datasets in Spark/Hadoop. The following items are covered during the course.

- Overview of the big data ecosystem for data scientists
- Linux system commands for data scientists
- Work with the hadoop file system (read/write, directories). Typical Hadoop files
- Map-Reduce & mapply
- Spark & R:
  - SparkSQL using package sparklyr & dplyr
  - Spark Machine Learning using package sparklyr (data preparation, regression, randomforest & boosted trees) + spark extensions
- HAWQ & R
  - Running PL/R stored procedures using Apache HAWQ
  - Classification & regression using MadLib & PivotalR

Analytical module V: Computer Vision with R and Python

Course duration: **2 day**

This course is a hands-on course covering the use of image analysis. It covers basic image manipulation, feature engineering techniques and finding patterns in images. The following items are covered.

- image manipulation & adjustments
- finding blobs, corners, gradients, edges & lines
- feature & object detection
- applying filters
- deep learning for image analysis
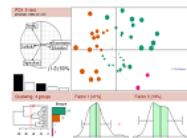- image segmentation

This course is offered also through LStat (https://lstat.kuleuven.be/training/index.htm).

Courses on the use of Oracle R Enterprise

About Oracle R Enterprise

Oracle has built "Oracle R Enterprise" as part of the Oracle Advanced Analytics Option to make the Open-Source statistical programming language R ready for the Enterprise and Big Data. Designed for problems involving large amounts of data, Oracle R Enterprise integrates R with Oracle Database. R users can run R commands and scripts for statistical and graphical analyses on data stored in the Oracle Database.

Oracle R Enterprise is a module of the Oracle Database 'Advanced Analytics' suite and consists of an R distribution which is supported by Oracle, an open source package ROracle to connect from R to Oracle, a set of packages called ORE allowing to transparently run R code inside, in parallel and embedded in the Oracle database (OREbase, OREcommon, OREdm, OREeda, OREembed, OREgraphics, OREmodels, OREpredict, OREstats, ORExml) and a set of data mining algorithms running inside Oracle and accessible from R.

More information about Oracle R Enterprise (ORE) can be found at the Oracle website: www.oracle.com/technetwork/database/database-technologies/r/r-enterprise/overview/index.html

The course is given in cooperation with our Oracle Partner Tripwire Solutions. You will be given access to an Oracle database where ORE is installed so you can use the ORE suite during the exercises.

ORE module I: ROracle and Oracle R Enterprise - transparancy layer

Course duration: **1 day**

In this session you will learn how to interface R with Oracle and use the transparency layer provided by ORE. This covers the following topics.

- ROracle - getting and sending SQL queries from Oracle
- Installing Oracle R Enterprise (ORE)
- Basic database connectivity: ore.exec, ore.ls, ore.synch, ore.push, ore.pull, ore.create, ore.drop, ore.get
- ORE data types: ore.character, ore.factor, ore.logical, ore.number, ore.datetime, ore.numeric. Conversion between data types
- ORE data structures: ore.matrix, ore.frame, ore.vector
- ORE transparancy data operations on ore.frame/ore.vector (subset, ncol, nrow, head, ifelse, paste, is.na, sd, mean, tapply, by, c, %in%, ...) and indexing and overwriting in-database ore.vectors
- Save R objects in Oracle ore.save, ore.load, ore.datastore and ORE data store handling
- Basic statistics with ORE (ore.univariate, ore.summary, ore.crosstab, ore.corr, exponential smoothing, t.test, wilcoxon, IQR)

You will be given access to an Oracle database where ORE is installed so you can use the ORE suite during the exercises.

ORE module II: Oracle R Enterprise - advanced data manipulation

Course duration: **0.5 days**

In this module, you will use the full power of ORE by running the ore group of functions which allow embedded R execution inside the Oracle database. This allows you to use the full power of Oracle and R together.
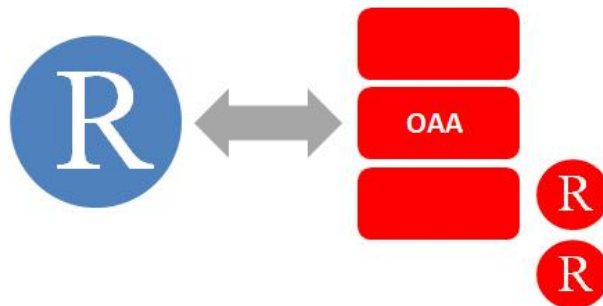
The session covers the following topics:

- Running R functions parallel inside the database: ore.doEval, ore.groupApply, ore.indexApply, ore.rowApply, ore.tableApply
- Creating R scripts inside the database and accessing ORE stored procedures
- Embedding R scripts in production database applications
- Embedded (parallel) R execution within ORE using the R Interface as well as the SQL Interface

You will be given access to an Oracle database where ORE is installed so you can use the ORE suite during the exercises.

ORE module III: Data mining models inside Oracle R Enterprise and Oracle Data Mining

Course duration: **1 day**

Data Mining (aka Machine Learning) refers to a set of statistical and mathematical techniques to reveal relationships and patterns in data. In contrast to 'classical' statistical methods, there is no need for hypothesis in advance. Applications of Data Mining Methods are forecast-models, market basket analysis, target group analysis and more.

In this session you will become acquainted with some of the most common data mining methods and learn how to use these algorithms in ORE. The following algorithms will be covered.

- principal component analysis and factor analysis
- kmeans clustering and orthogonal partitioning
- data reduction using Minimum Description Length attribute importance
- linear models and generalized linear models
- naive bayes, neural networks, decision tree and support vector machines
- market basket analysis / recommendation engines (apriori)
- bagging

You will be given access to an Oracle database where ORE is installed so you can use the ORE suite during the exercises.

Mark that if you are unfamiliar with data science algorithms, it is advised to also follow the module on 'Statistical machine learning with R'