Applied Statistics and Data Analysis

1. Introduction

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

Thanks to R. Bellio and L. Grassetti for some useful hints

Table of contents

General information

- **Timetable:** Monday 12:30-14:30, Wednesday 14:30-16:30 (room A029 LAB2 DMIF).
- **Computer Lab:** 12 hours out of 48 (students may use personal laptops).
- **Office hours:**
  - Wednesday 09:00-11:00 at Dipartimento di Scienze Economiche e Statistiche, via Tomadini 30/a or using Teams;
  - for further office hours, please contact me at classes or by e-mail.
- **Teaching material:** available at `https://elearning.uniud.it/moodle/`.
- **Prerequisites:** a first course in probability and statistics.

  For a **review on statistics** (or a first introduction) consider the teaching material of the course of *Statistics* (*Laurea TWM/IBW/IBML*), with regard to descriptive and inferential statistics, and the recordings of the 2020/2021 Lecture 3, both available at `https://elearning.uniud.it/moodle/`.

Course Outline

- **Introduction to statistics and data analysis**
- **Exploratory data analysis**
- **A review of inference concepts** (*for individual revising*)
- **Linear regression with a single predictor**
- **Towards multiple linear regression and logistic regression**
- **Predictive and classification methods**
- **Unsupervised methods (principal component analysis, cluster analysis)**
- **Tree-based methods**

Implementation of methods will be shown in the labs using the R software (`https://www.r-project.org/`).

What is statistics?

(Inspired by a publication of the American Statistical Association www.amstat.org/careers/whatisstatistics.cfm)

- "Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty" (Davidian, M. and Louis, T. A., 10.1126/science.1218685).
- Statistical methods are developed in a synergy with the relevant supporting mathematical theory, and more recently with computing, and they are applied in a wide variety of scientific, social, and business frameworks.
- Statistical methods provide crucial guidance in determining what information is reliable and which predictions can be trusted.
- Statistical methods often help search for clues to the solution of a scientific mystery and sometimes keep investigators from being misled by false impressions.

What do statisticians do?

- The world is becoming "quantitative" and more and more professions depend on data and numerical reasoning.
- Statisticians are experts in producing trustworthy data, analyzing data to make their meaning clear and drawing practical conclusions from data.
- Statisticians work with people from other professions to solve practical problems and they must know more than statistics.
- "The best thing about being a statistician is that you get to play in everyone else's backyard" (John Tukey, Bell Labs, Princeton University).

  Biography of John M. Tukey
  `https://magazine.amstat.org/blog/2001/06/16/tukey-sih/`

The Joy of Stats (BBC documentary, 2010)
`https://www.gapminder.org/videos/the-joy-of-stats/`

Dream job of the next decade?

https://www.youtube.com/watch?v=pi472Mi3VLw

Hal Varian

Economist

Hal Ronald Varian is an economist specializing in microeconomics and information economics. He is the Chief Economist at Google and he holds the title of emeritus professor at the University of California, ...

Wikipedia

Born: March 18, 1947 (age 66), Wooster, Ohio, United States

Education: University of California, Berkeley (1973), Massachusetts Institute of Technology (1969)

New computing tools

(Inspired by the Preface of *Data Analysis and Graphics Using R*)

- The recent advances in statistical computing methodology have made possible the development of new powerful tools for data analysis and prediction.
- New types of data and data sets of unprecedented size (e.g. textual data, image data), combined with new data analysis demands, boost the development of new hybrid data analysis approaches, such as machine learning, data mining and analytics.
- However, the traditional concerns of professional data analysts remain as important as ever. The size and the complexity of data set are not itself a guarantee of quality and of relevance to issues that are under investigation.
- No amount of statistical or computing technology can be a substitute for skill in the use of statistical analysis methodology. Statistical software systems are one of several components of effective data analysis.

The main steps of a statistical analysis

- Formulation of the problem: understand the background, specify the objectives of the analysis, put the problem into statistical terms.
- Collection and organization of data: observational or experimental data, missing values, units of measurements, codification of data, organization of data.
- Initial data analysis: numerical and graphical summaries to get into data.
- Data analysis.
- Presentation of the results.

"The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill" (Albert Einstein)

Statistics as a way of life

Statistics and Machine Learning fill different places in the large ecosystem of *Data Science*.

Statistics and ML share several techniques, but there are some **crucial differences** between the two disciplines.

In our view, some notable points of the *Statistics way* are as follows:

- The importance of context;
- The role of principles;
- Statistical models, aka making sense of data;
- A lingua franca.

The importance of context

- Statistics is always concerned with the **data context**, so that data analyses are never fully automatic. A preliminary understanding of the way data are produced is essential for any further step.
- Solutions can be **adapted** to a different setting with the necessary changes, but only after studying the context, and never automatically.
- This approach is less fruitful or even misplaced for some tasks which can be solved by a purely algorithm approach, such as some pattern recognition tasks.

The role of principles

- **Statistical principles** are of paramount importance. They range from optimality in estimation/prediction to principles adhering to likelihood theory and Bayesian theory.
- Sometimes they are inspired by highly-stylized settings, but they provide guidelines which are useful also in intricate real-life scenarios.

Statistical models, aka making sense of data

- **Statistical models** are of central importance. They are mathematical description of the *data generating process*, and they include both deterministic and random components.
- Even methods which are strongly algorithmic in nature (e.g. regression trees or LDA in text mining) are often interpreted as based on statistical models, with the possibility of applying *general statistical principles*, such as model selection criteria or Bayesian inferential techniques.
- Statisticians usually endorse the **Occam's razor principle**, avoiding overly complex models unless the data available are large/rich enough. (That's why we are not so fond of neural networks, which are sometimes misused nowadays).

A lingua franca

- Statisticians, with few exceptions, have adopted the open source R software.
- R is both a statistical software as well as a programming language, with interfaces to many data mining/ML software, such as Weka and H2O. (Despite what many people outside statistics believe, *R is far more powerful and versatile than Matlab, it's not just a free version of it*).
- Having a common language has simplified things a lot within the Stat community.

Table of contents

Business analytics

- Business analytics focuses on understanding business performance and on developing new insights based on data analysis.
- Nowadays, more and more companies collect data, sometimes in enormous amounts, and a suitable effective usage of these data can be crucial for business.
- Business analytics makes extensive use of statistical procedures, including descriptive and inferential analysis and predictive modeling, to describe the main feature of business processes and to drive decision making.
- "Data-driven" companies consider their data as a corporate asset, crucial for competitive advantage. "Patterns emerge before the reasons for them become apparent" (Vasant Dhar, Stern School of Business, New York University).
- Business analytics explores data to find patterns and relationships, to explain the occurrence of certain results and to forecast future scenarios.

Business analytics

From Wikipedia, the free encyclopedia

Not to be confused with Business analysis.

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(October 2010)*

Business analytics (**BA**) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.[1] Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods.

Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling,[2] and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (that is, predict), what is the best that can happen (that is, optimize).[3]

Examples of application [edit]

Banks, such as Capital One, use data analysis (or *analytics*, as it is also called in the business setting), to differentiate among customers based on credit risk, usage and other characteristics and then to match customer characteristics with appropriate product offerings. Harrah's, the gaming firm, uses analytics in its customer loyalty programs. E & J Gallo Winery quantitatively analyzes and predicts the appeal of its wines. Between 2002 and 2005, Deere & Company saved more than $1 billion by employing a new analytical tool to better optimize inventory.[3] Example : It can help you focus on the fundamental objectives of the business and the ways analytics can serve them. A telecoms company that pursues efficient call centre usage over customer service might save money.

Types of analytics [edit]

- Decisive analytics: supports human decisions with visual analytics the user models to reflect reasoning.[4]
- Descriptive Analytics: Gain insight from historical data with reporting, scorecards, clustering etc.
- Predictive analytics (predictive modeling using statistical and machine learning techniques)
- Prescriptive analytics recommend decisions using optimization, simulation etc.

Business and (big) data

February 11, 2012

The Age of Big Data

By STEVE LOHR

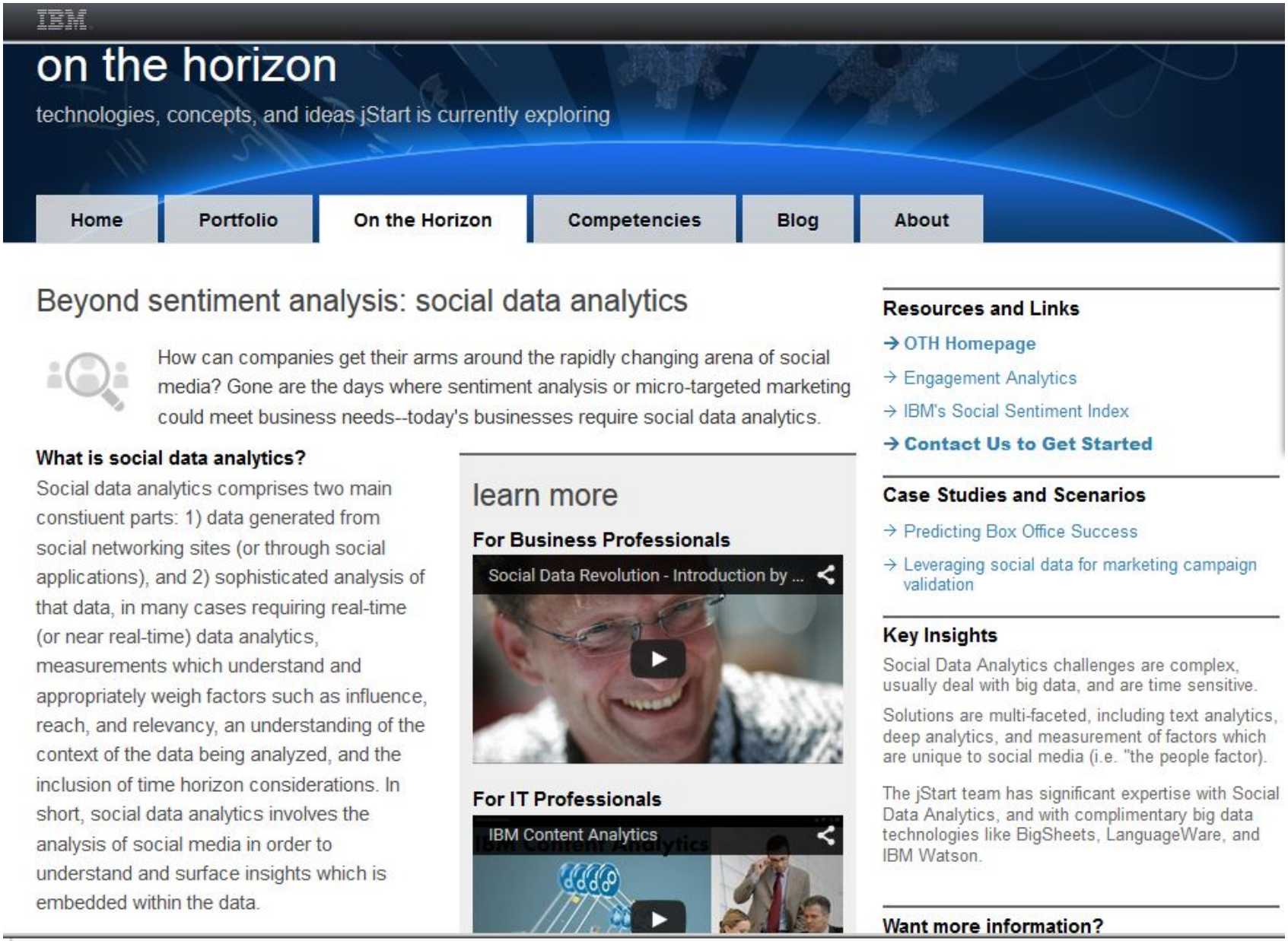
GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

Welcome to the Age of Big Data. The new megarich of Silicon Valley, first at Google and now Facebook, are masters at harnessing the data of the Web — online searches, posts and messages — with Internet advertising. At the World Economic Forum last month in Davos, Switzerland, Big Data was a marquee topic. A report by the forum, "Big Data, Big Impact," declared data a new class of economic asset, like currency or gold.

Social data analytics

- Social data analytics is the practice of gathering and analyzing social data to learn about human behavior.
- Social data corresponds to information generated by a variety of human activities, including social media and internet use. Social data are often "big data", usually too large and varied to be analyzed with ordinary statistical tools.
- Social data are not only important for business, but totally necessary for social investigations such as opinion surveys, market analyses, public health surveillance, etc.
- Social data analysis is challenging since it involves a number of factors (e.g. context, content, sentiment), it is time-sensitive and it usually assumes spatial and network dependencies.
- Social data analytics includes sentiment analysis, and in particular costumer sentiment analysis, text mining and social network analysis.


IBM
on the horizon
technologies, concepts, and ideas jStart is currently exploring
Home
Portfolio
On the Horizon
Competencies
Blog
About
Beyond sentiment analysis: social data analytics
How can companies get their arms around the rapidly changing arena of social media? Gone are the days where sentiment analysis or micro-targeted marketing could meet business needs--today's businesses require social data analytics.
What is social data analytics?
Social data analytics comprises two main constiuent parts: 1) data generated from social networking sites (or through social applications), and 2) sophisticated analysis of that data, in many cases requiring real-time (or near real-time) data analytics, measurements which understand and appropriately weigh factors such as influence, reach, and relevancy, an understanding of the context of the data being analyzed, and the inclusion of time horizon considerations. In short, social data analytics involves the analysis of social media in order to understand and surface insights which is embedded within the data.
learn more
For Business Professionals
Social Data Revolution - Introduction by ...
For IT Professionals
IBM Content Analytics
Resources and Links
→ OTH Homepage
→ Engagement Analytics
→ IBM's Social Sentiment Index
→ Contact Us to Get Started
Case Studies and Scenarios
→ Predicting Box Office Success
→ Leveraging social data for marketing campaign validation
Key Insights
Social Data Analytics challenges are complex, usually deal with big data, and are time sensitive.
Solutions are multi-faceted, including text analytics, deep analytics, and measurement of factors which are unique to social media (i.e. "the people factor).
The jStart team has significant expertise with Social Data Analytics, and with complimentary big data technologies like BigSheets, LanguageWare, and IBM Watson.
Want more information?

IBM

Current examples of social data analytics

Social data analytics exists today: when the Annenberg Innovation Lab decided to track and understand how sentiment evolved and impacted the Arab Spring movements sweeping the Middle East, they employed social data analytics. Further, when the same lab decided to understand how sentiment might be able to predict box office potential for yet-to-be-released movies, they also performed social data analytics. When one of the world's leading research institutions wanted to understand how social media could impact how their clients conducted public awareness campaigns, social data analytics also played a role. Most companies understand that social media may have important roles to play in how they conduct business...the problem is that most are unaware of how to go about tackling the problem.

Solutions exist today

jStart has been leading exploration of this space by combining the skills of the jStart team with regards to sophisticated data analytics, text analytics, and specific tooling:

Key concepts to understand in social data analytics

When talking about social data analytics, there are a number of factors it's important to keep in mind (which we noted earlier):

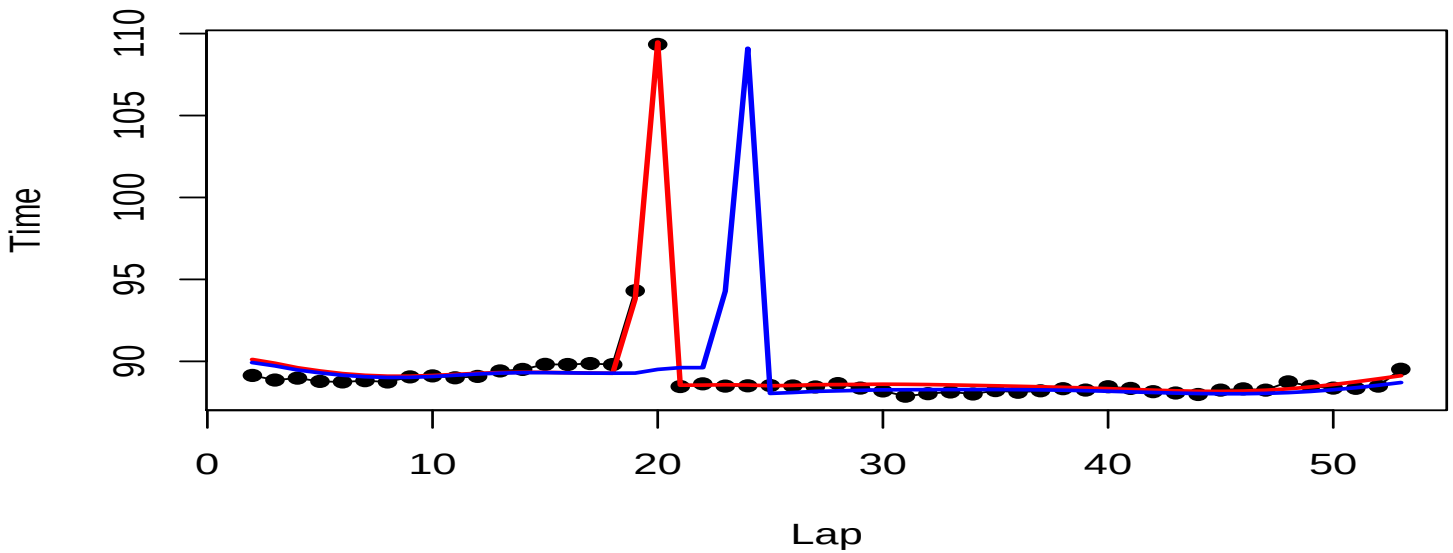
- **Sophisticated Data Analysis:** what distinguishes social data analytics from sentiment analysis is the depth of the analysis. Social data analysis takes into consideration a number of factors (context, content, sentiment) to provide additional insight.
- **Time consideration:** windows of opportunity are significantly limited in the field of social networking. What's relevant one day (or even one hour) may not be the next. Being able to quickly execute and analyze the data is an imperative.
- **Influence Analysis:** understanding the potential impact of specific individuals can be key in understanding how messages might be resonating. It's not just about quantity, it's also very much about quality.
- **Network Analysis:** social data is also interesting in that it migrates, grows (or dies) based on how the data is propagated throughout the network. It's how viral activity starts--and spreads.

Want more i
Want to know
Contact us

Table of contents

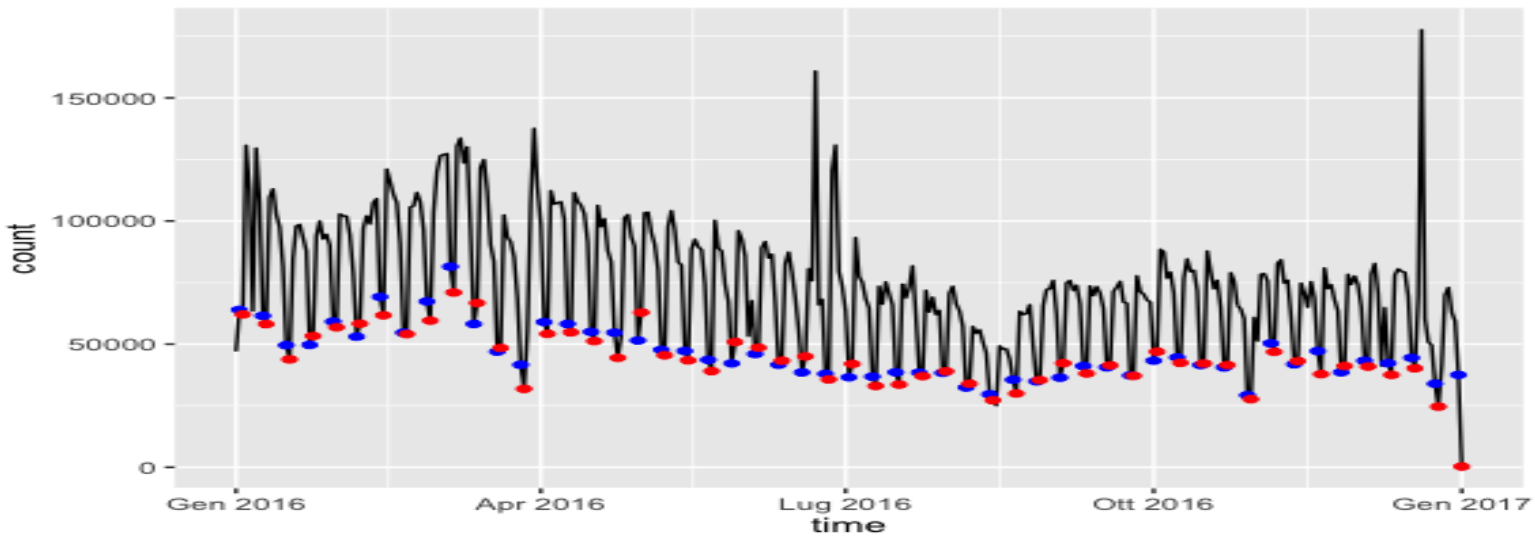
Formula 1 lap time modeling

- The aim is to study **Formula 1 driver performance**, focusing on **lap time evolution** described as a function of the **explanatory variables** Driver, Team, Pit Stop, Fuel, Tyres and Traffic.
- Formula 1 season 2015, Italian grand prix: optimal pit stop strategy for Massa (Williams).
- **Observed lap times** and lap times prediction with a pit stop at the 19th lap (the lap chosen by the team) and at the 23rd lap (the lap indicated by the model): more than two seconds gained.



Cybersecurity: analysis of log files

- Data set with apache log files: text files which give information on "what happened when by whom" with regard to the use of the web applications of a company.
- **Anomaly** is a pattern in the data that does not conform to the expected behavior: it can be determined by a cyber attack.
- Statistical procedures which are able to rapidly detect when anomalies are occurring, in time for preventive action to be taken.
- Count of daily log files from January 2016 to January 2017: two anomalies (end of June and end of December)



Predictive maintenance

- Monitoring of a **Quench Tower**, which is a machinery where a stream of gas is rapidly cooled through a liquid (typically water).
- Time series data which come from sensors that have the function of monitoring the flow and regulating the water level by means of control valves.
- Five main time series called LEVEL1, LEVEL2, VALVE1, VALVE2 e FLOWQT.
- The aim is to monitor the time series to identify **regime changes**, related to the mean level and to the variability: possible anomalies that require a maintenance of the plant.

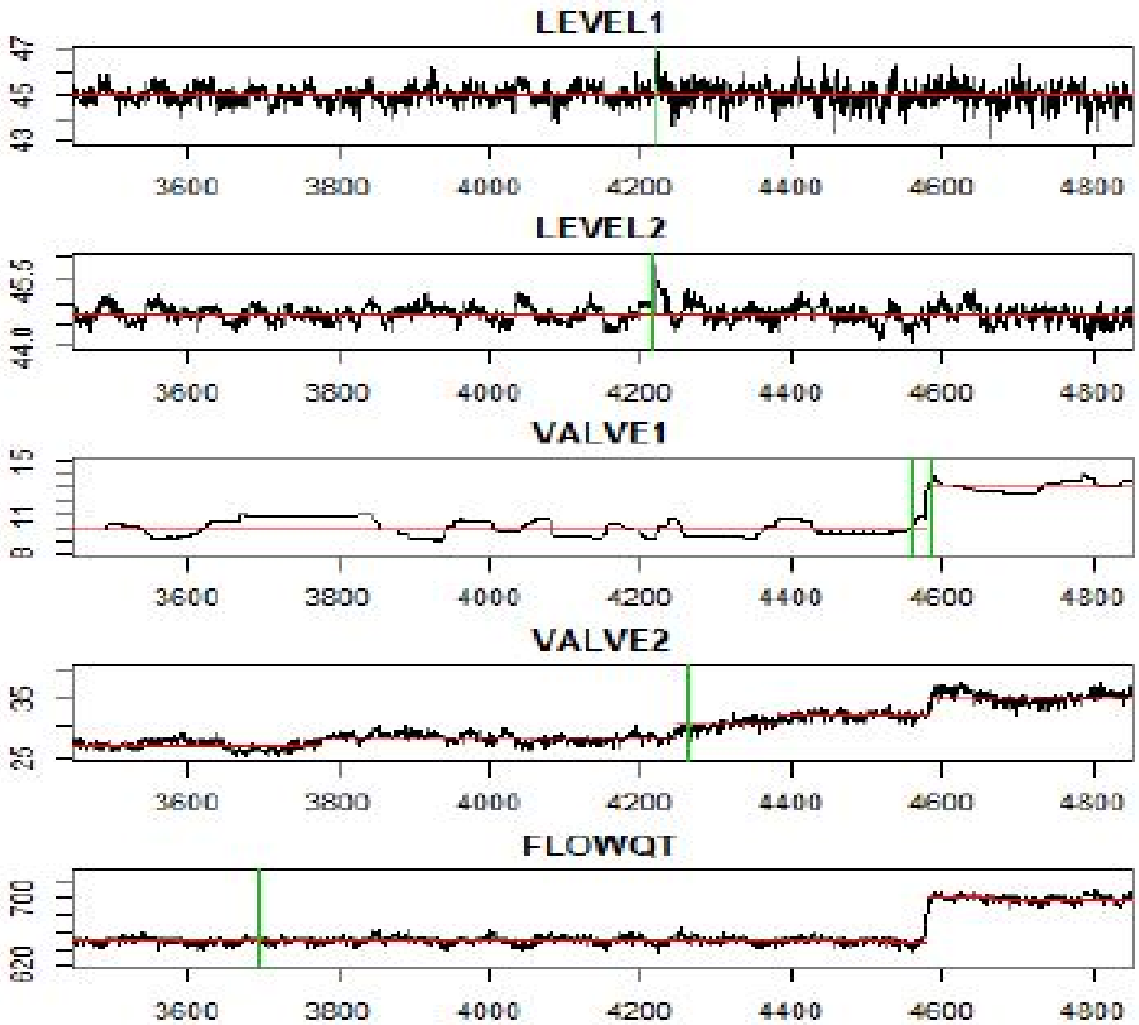

LEVEL1
43 45 47
3600 3800 4000 4200 4400 4600 4800
LEVEL2
44.0 45.5
3600 3800 4000 4200 4400 4600 4800
VALVE1
8 11 15
3600 3800 4000 4200 4400 4600 4800
VALVE2
25 35
3600 3800 4000 4200 4400 4600 4800
FLOWQT
620 700
3600 3800 4000 4200 4400 4600 4800

Table of contents

Course nature

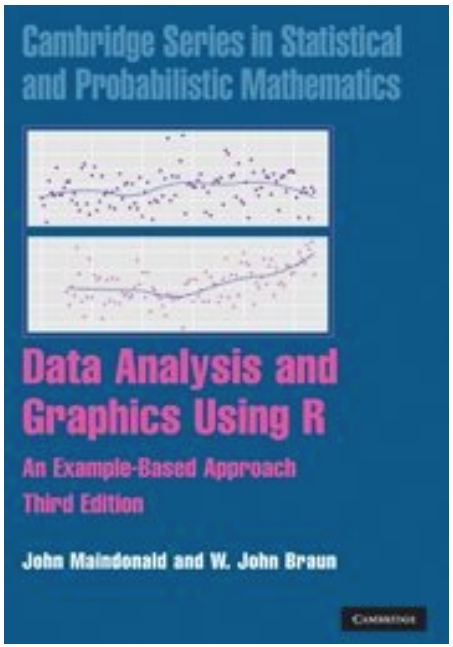
- The course focuses on statistical methods, with emphasis on applied aspects. It reviews and introduces some basic statistical procedures and some elementary statistical models.
- The course gives the essential background for attending advanced courses on statistics and statistical learning.
- It starts essentially from scratch, assuming no prior background on probability and statistics. However, basic concepts will be presented rather cursorily, so that some familiarity with probability and statistics is recommended.
- Of fundamental importance will be the basic elements of descriptive statistics and statistical inference (random sampling, statistical model, point estimation, confidence interval, statistical test).
- The main ideas and the fundamental principles underlying statistical methods are presented also through examples, without emphasizing the mathematical details, according to the principle of *learning (and reviewing) by doing*.

Suggested (supplementary) text

J. Maindonald and W.J. Braun: *Data Analysis and Graphics Using R - An Example-Based Approach* (Third Edition). Cambridge University Press, 2010.

The book has a practical nature with focus on applications, with little prior knowledge required.

There is an associated webpage with a lot of useful adds-on:
`https://maths-people.anu.edu.au/~johnm/r-book/daagur3.html`



Additional references

- P. Dalgaard: *Introductory Statistics with R*. Springer, 2008.
- J. Ledolter and R.V. Hogg: *Applied Statistics for Engineers and Physical Scientist* (Third Edition). Prentice Hall, 2009.
- J.P. Marques de Sá: *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. Springer, 2007.
- G. James, D. Witten, T. Hastie and R. Tibshirani: *An Introduction to Statistical Learning*. Springer, 2013. The new edition (and the old one) is freely available at `https://www.statlearning.com/`.
- OpenIntro Statistics, which is a free textbook available at `https://www.openintro.org/`.

  A lot of additional teaching and learning facilities may be find: labs for R, videos, forums, data sets, additional textbooks, the R package `openintro`, containing data and tools used in the textbook.

Table of contents

The R project for statistical computing

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Mailing Lists
Bug Tracking
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- Beta test period for version 3.3.0 has been extended to accommodate new Windows toolchain for CRAN. Final release rescheduled for Tuesday 2016-05-03.
- **Notice XQuartz users (Mac OS X)** A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- **R version 3.2.4 (Very Secure Dishes)** has been released on Thursday 2016-03-10.
- **R version 3.3.0 (Supposedly Educational) prerelease versions** will appear starting Monday 2016-03-14. Final release is scheduled for Thursday 2016-04-14.
- The **R Logo** is available for download in high-resolution PNG or SVG formats.
- **useR! 2016**, will take place at Stanford University, CA, USA, June 27 - June 30, 2016.
- **The R Journal Volume 7/2** is available.
- **R version 3.2.3 (Wooden Christmas-Tree)** has been released on 2015-12-10.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.

Wikipedia entry

R (programming language)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsource challenged and removed. *(January 2016)*

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.[3] The R language is widely used among statisticians and data miners for developing statistical software[4] and data analysis.[5] Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.[6]

R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme.[7] S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered.[8]

R was created by Ross Ihaka and Robert Gentleman[9] at the University of Auckland, New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S.[10]

R is a GNU project.[11] The source code for the R software environment is written primarily in C, Fortran, and R.[12] R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.[13]

Contents [hide]

The R statistical software

- The theoretical part of the course is fully integrated with the associated lab part, based on applications using the R statistical software.
- R is a free software environment for statistical computing and graphics (`https://www.r-project.org/`).
- The R project was started in the middle of the 1990s by two researchers working in New Zealand, **R**obert Gentleman and **R**oss Ihaka.
- R is an open source project based on the S programming language. It is actually one of the available implementations of S (the other one is the commercial S-PLUS software).
- R studio (`https://www.rstudio.com/`) is an integrated development environment for R. It is a powerful user interface for R, including a console, an editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

  R Studio is also available in a free, open source edition.

Development of R

- R is currently being developed by a group researchers forming the *R Core Group*, but also from many volunteers.

  The number of daily users of R is something on the order of several hundred thousands (at least), by far more than any other competitor such as SAS or IBM SPSS.
- There are several thousands of *R packages*, freely provided by contributors for applications in finance, economics, computer science, biology, medicine, sociology, and more.
- Such immense richness of resources is invaluable, making R *de facto* the most important statistical software for both academic research and (potentially) business applications.

  In the USA, R is currently used by companies such as *Google*, *Pfizer*, *Bank of America*.
- We use R for several reasons: versatility, interactivity, freedom, popularity.

Basic features of R

- Environment developed for representing **data and statistical models**.
- Powerful **graphical** capabilities.
- Based on **object-oriented programming language**, can be easily extended by users.
- Free and **open source**: users can access the code and modify it freely.
- **Multi-platform**, namely it runs on all the existing operating systems for desktop/laptop computers, with some options for tablets and smartphones.
- Can be used for **huge data sets**, with simple interface with main **database** software.
- Can be interfaced with **other programming languages**, such as C, C++, Fortran and Java.

Documentation for R

- Besides the main webpage, there are archives (e.g. the CRAN, from which the software can be downloaded, `https://cran.r-project.org/`), mailing lists, forums, blogs, GitHub resources, etc.
- A lot of documentation available, both on CRAN and on the web.
- Many bibliographic references. Among the others, we mention:

  Iacus, S. M. e Masarotto, G. (2007). *Laboratorio di statistica con R (Seconda edizione)*. McGraw-Hill.

  Ieva, F., Masci, C. e Paganoni, A.M. (2016). *Laboratorio di statistica con R (Seconda edizione)*. Pearson.

  Wickham, H. and Grolemund, G. (2017). *R for Data Science*. O'Reilly. (`https://r4ds.had.co.nz/`).

  Long, J.D, and Teetor, P. (2011). *R Cookbook*, 2nd Ed., O'Reilly. (`https://rc2e.com/`).

  The book by Maindonald and Braun and the book by Dalgaard provide a good introductory documentation of the software.

R in this course

- The R software is an essential tool for this course. It will be employed to apply all the methods introduced during classes.
- Will be used in the computer labs, but sometimes also in lectures.
- The intensive use of software will make the course to be strictly adhering to the principle of *learning by doing*.
- Students are expected to learn by themselves the introductory notions on R (main features, session management, commands, data structures, reading data, exporting data, graphics, writing functions).
- A non-superficial knowledge of the R software will be essential for a successful final exam.

Table of contents

Information on the final exam

- The final exam consists of a written part and of an oral part (only for students who passed the written part).
- The **written test** covers the entire course, including the R software: **score up to 26 points** (minimum score 15 points).

  The **oral presentation** is on a topic assigned at the end of the course: **score up to 7 points**.
- The assignment will be given to each student or teams of two students.

  The required work consists of two parts:

  - writing an **original** report (*approximately 15/20 pages*);

  - an oral presentation (*up to 15 minutes*) focusing on the results included in the report.

More on the written report

- The oral exam takes place **in the same session** (*appello*) of the written exam. There is only one exception: if one member of the team takes the oral exam, also the other members will take it, even if they haven't passed the written test yet.

  Namely, the oral presentation will be always made in a single occasion.
- The written report has to be delivered to the instructor in printed or electronic form **at least five days before the presentation**.
- The report is about an R package or about a relevant topic, not encountered during the course.
- The requirement is to study the relevant theory, learn how to use it and illustrate its usage in some applications.
- The presentation should follow the report, with less emphasis on the software details and more on the practical usage of the package.

Homework

- **Watch** *The Joy of Stats* (BBC documentary, 2010) `https://www.gapminder.org/videos/the-joy-of-stats/`.
- **Read** the slides prepared for the course of *Statistics*, *Laurea TWM/IBW/IBML* and watch the recordings of the 2020/2021 Lecture 3.
- **Download and install** R and R studio.
- **Begin to use** R, using the documentation provided for Lab 1 or the documentation available from the CRAN and around the web.