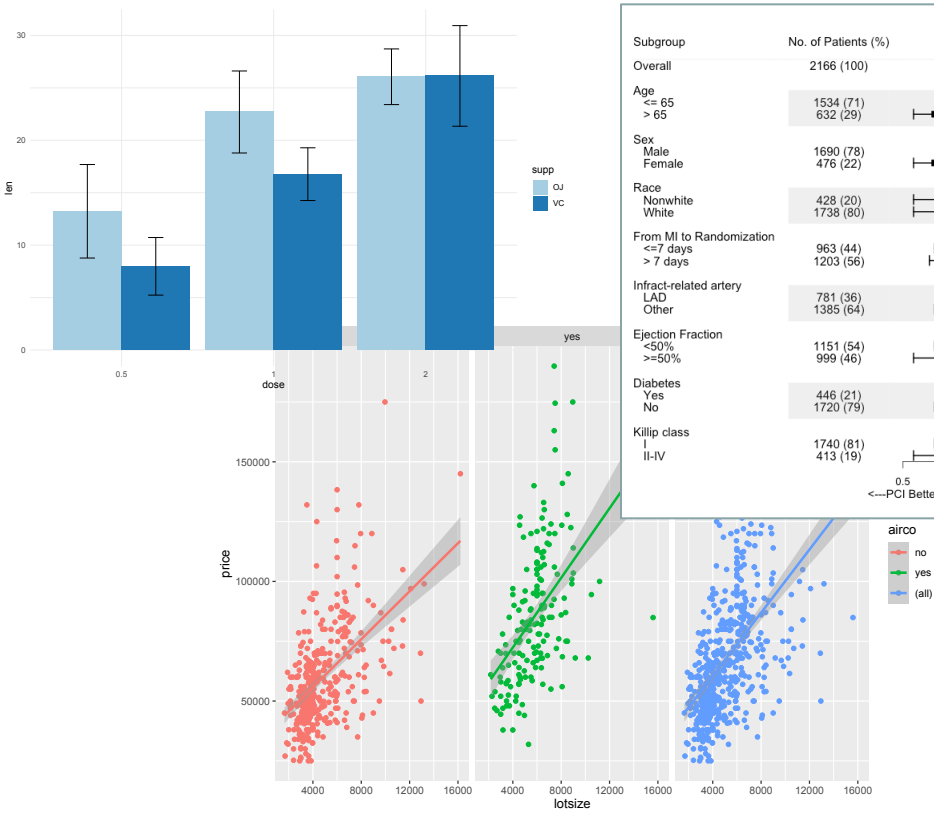
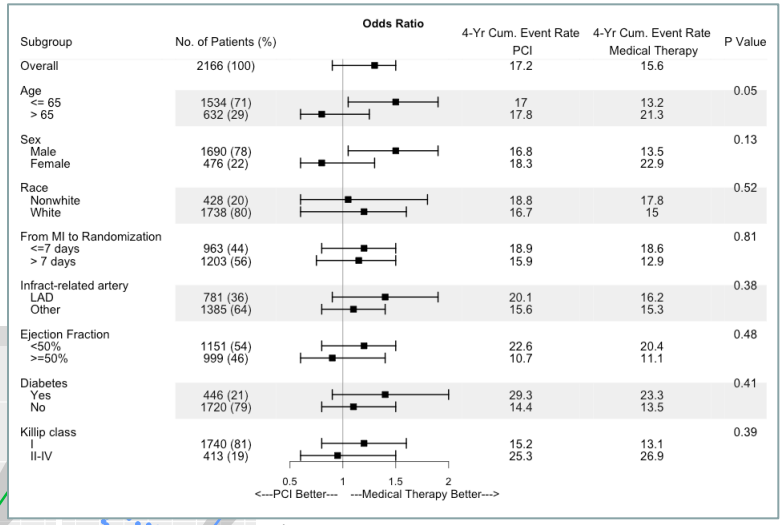
Instituto de Salud Carlos III

Introducción a R





Pablo Fernández Navarro

Grupo de Bioinformática y Gestión de Datos (BIODAMA)
Unidad de Epidemiología del Cáncer y Ambiental
Centro Nacional de Epidemiología
Instituto de Salud Carlos III
Monforte de Lemos 5
28029 Madrid (Spain)
Tph: 918222644
Fax: 913877815
e-mail: pfernandezn@isciii.es

Mario González Sánchez

Grupo de Bioinformática y Gestión de Datos (BIODAMA)
Unidad de Epidemiología del Cáncer y Ambiental
Centro Nacional de Epidemiología
Instituto de Salud Carlos III
Monforte de Lemos 5
28029 Madrid (Spain)
Tph: 918222876
Fax: 913877815
e-mail: mariogonzalez@isciii.es

Instituto de Salud Carlos III

BIODAMA

Bioinformatics and Data Management

Instituto de Salud Carlos III

“El año que novicio fui, espantome;

quíseme retirar, pero no hay cosa

que el tiempo y la costumbre no la dome”

Arturo Pérez Reverte. *Corsarios de Levante*

Instituto de Salud Carlos III

1. Introducción

2. Librerías (packages)

3. Manejo básico de datos

    3.1. Introducción objetos en R
    3.2. Importación
    3.3. Data frame
    3.4. Exportación

4. Gráficos y Funciones

Instituto de Salud Carlos III

1. Introducción (I)

Plan de Formación Interna

"Nuevas Tecnologías"

Índice

1.1. Que es R
1.2. Carácterísticas
1.3. ¿Por qué R?
1.4. Enlaces
1.5. Libros y material general
1.6. Screenshots
1.7. Comunidad R
1.8. Actualidad

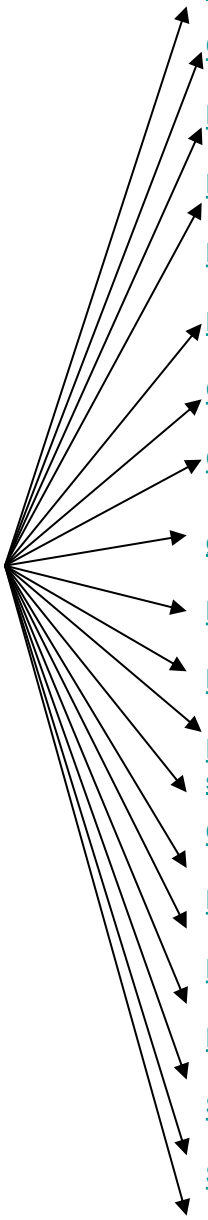
1.1. Qué es R?

- **R** es un lenguaje y entorno de programación para análisis estadístico y gráfico **gratuito** (“high-level language)
- Es un **lenguaje de programación orientado a objetos**.
- Proyecto de software libre, resultado de la implementación GNU del premiado lenguaje “S” y “Scheme”.
- R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística.
- Muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras.

1.2. Características (1/2)

- ***Versatilidad:*** posibilidad de cargar diferentes librerías o paquetes con finalidades específicas de cálculo o gráfico.
- El proyecto más conocido desarrollado sobre R es Bioconductor.
- Rmetrics está orientado al análisis de los mercados financieros y la valoración de instrumentos de inversión.
- R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.
- Algunas de las características más relevantes de R son que se trata de un lenguaje interpretado y que está orientado a objetos.

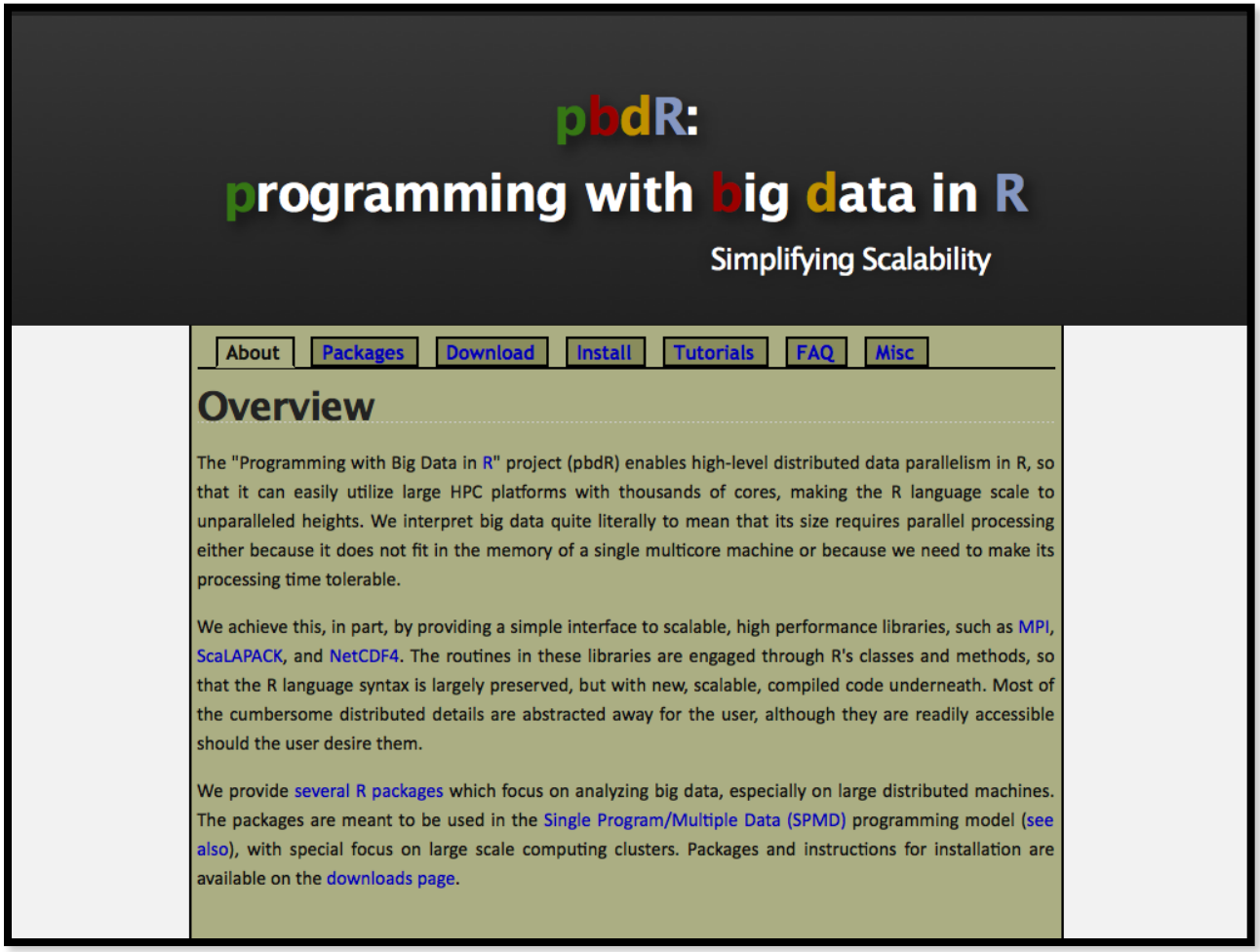
Instituto de Salud Carlos III



Versatilidad

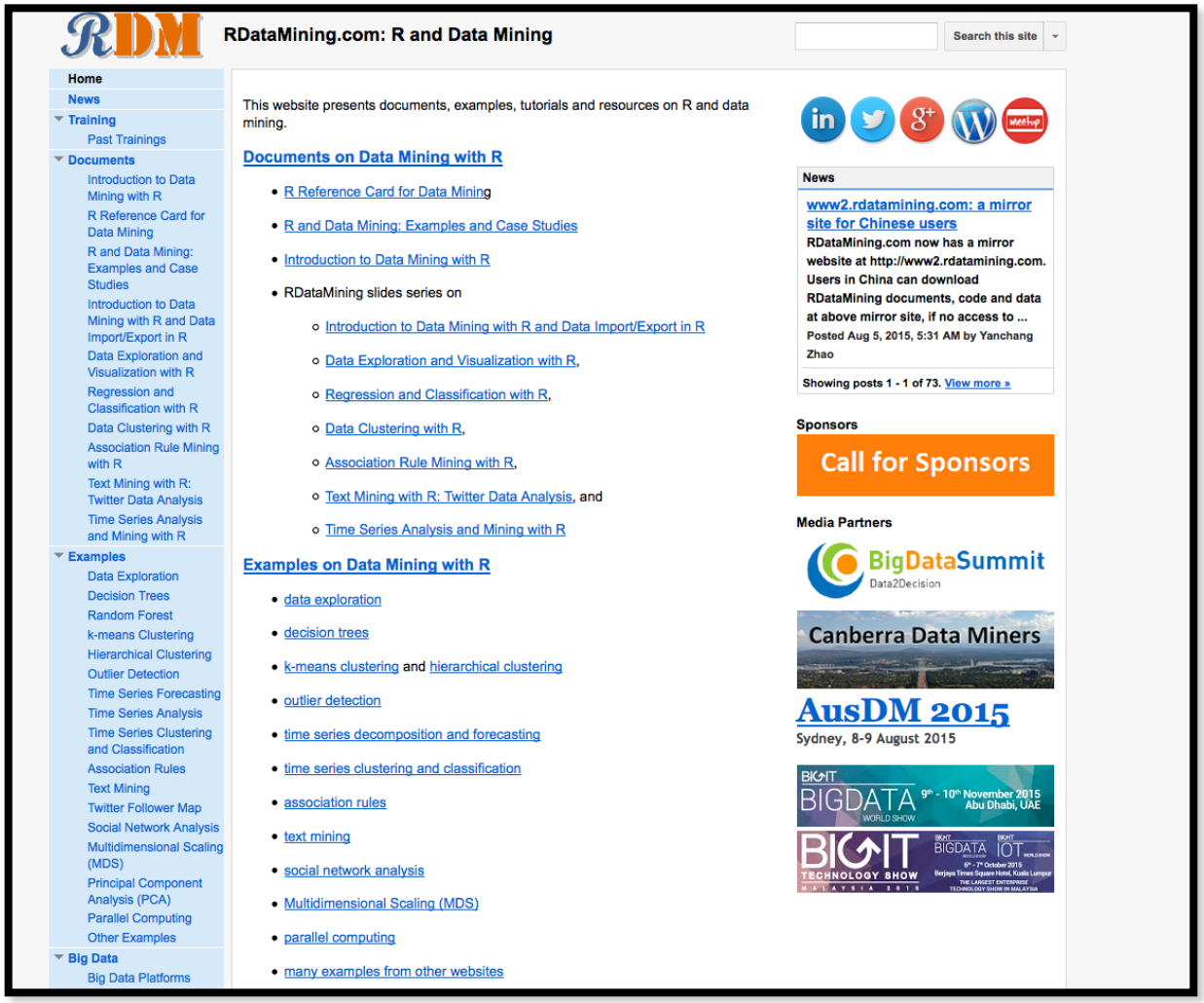
Bayesian	Bayesian Inference
Cluster	Cluster Analysis & Finite Mixture Models
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
gR	gRaphical Models in R
MachineLearning	Machine Learning & Statistical Learning
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Psychometrics	Psychometric Models and Methods
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
Survival	Survival Analysis

Instituto de Salud Carlos III



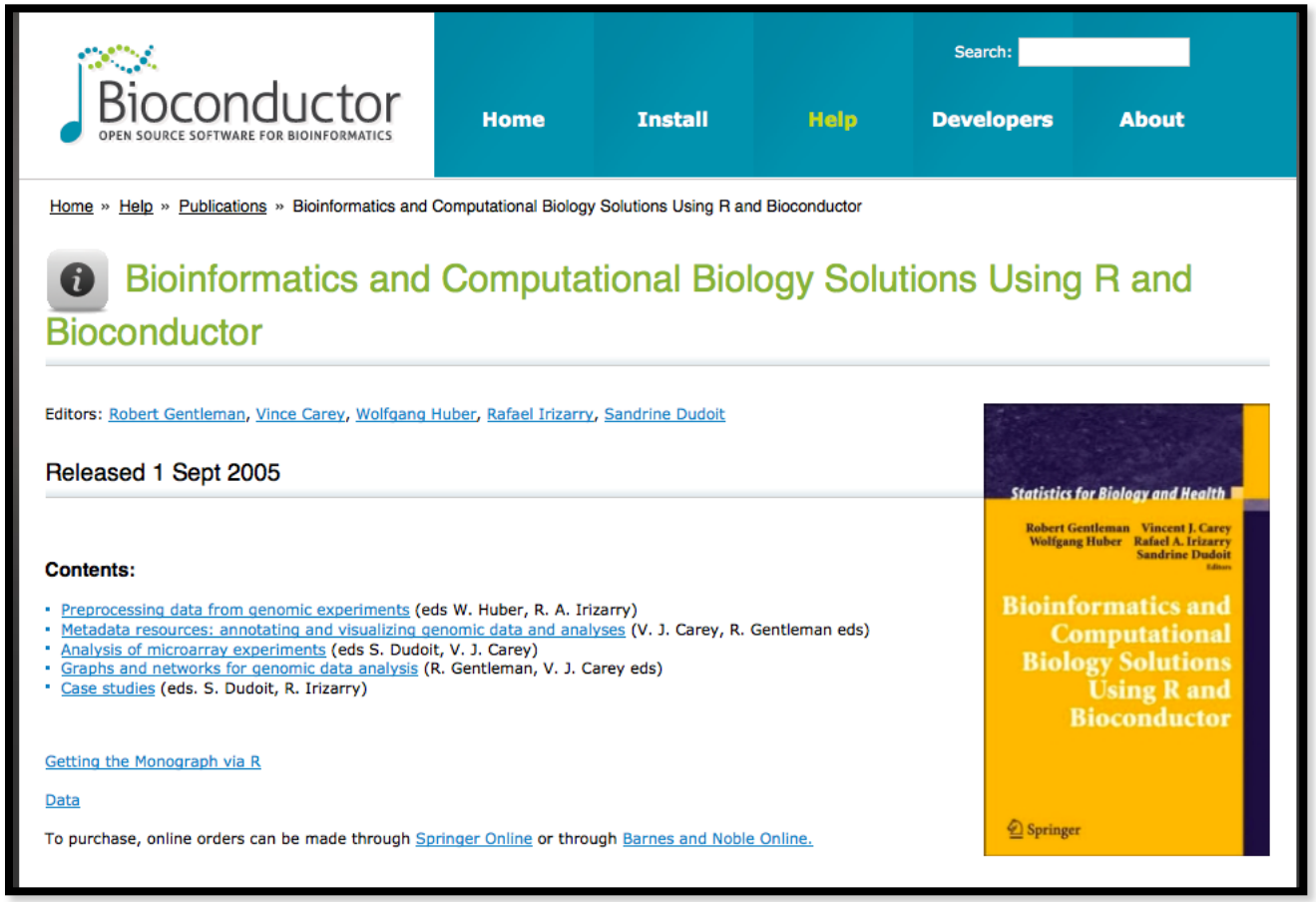
http://www.r-pbd.org/

Instituto de Salud Carlos III



http://www.rdatamining.com/

Instituto de Salud Carlos III



http://www.bioconductor.org/help/publications/books/bioinformatics-and-computational-biology-solutions/

1.2. Características (2/2)

- Puede integrarse con distintas bases de datos
- Existen librerías que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python o en lenguajes de código compilado, por ejemplo,C o Fortran.
- Una amplia colección de librerías se encuentran en CRAN.
- Existen diversas interfaces gráficas para R, como R-Commander.
- Se parece a Matlab y a Octave, y su sintaxis recuerda a C/C++.

1.3. ¿Por qué R? (1/2)

- **No hay razón en cambiar si se realizan tareas sencillas ya implementadas y no se intentará nada nuevo en el futuro**
- **Principal razón para cambiar a R**: *cobertura* de análisis y *disponibilidad de nuevas aplicaciones* en campos por ejemplo de modelos lineales generalizados.
- Ser capaz de entender la literatura.
- La mayoría de "top people" en las distintas disciplinas lo usa.

1.3. ¿Por qué R? (2/2)

- Una gran parte de los estadísticos lo utiliza.
- Gran soporte y seguro.
- Una gran red de respaldo.
- Capacidad de programar y construir funciones propias
- Es **GRATIS**

1.4. Enlaces (1/2)

- El proyecto R:

  http://www.r-project.org/

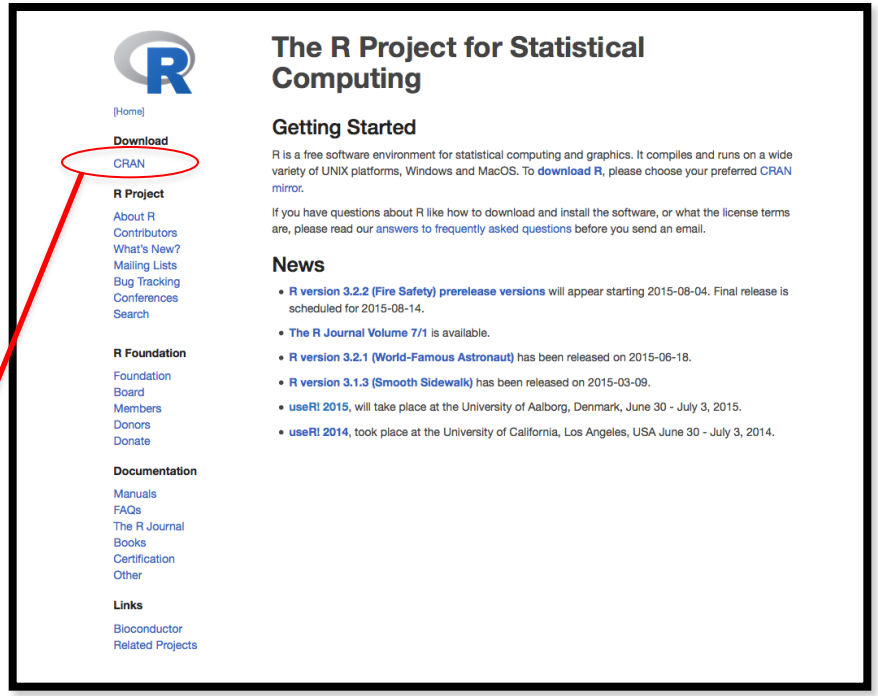
- The R Reference Manual - Base Package by the R Development Core Team. ISBN 0-9546120-0-0 (vol. 1), ISBN 0-9546120-1-9 (vol. 2) :

  http://www.network-theory.co.uk/R/base/

- Colección librerías en CRAN (Comprehensive R Archive Network):

  https://cran.r-project.org/

Instituto de Salud Carlos III



[Home]

Download

CRAN

R Project

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

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

Links

Bioconductor
Related Projects

The R Project for Statistical Computing

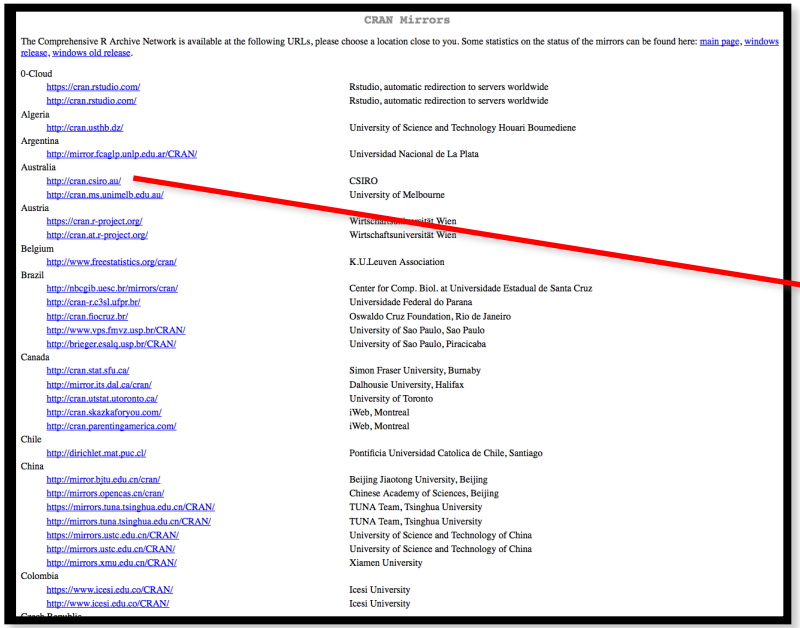
Getting Started

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

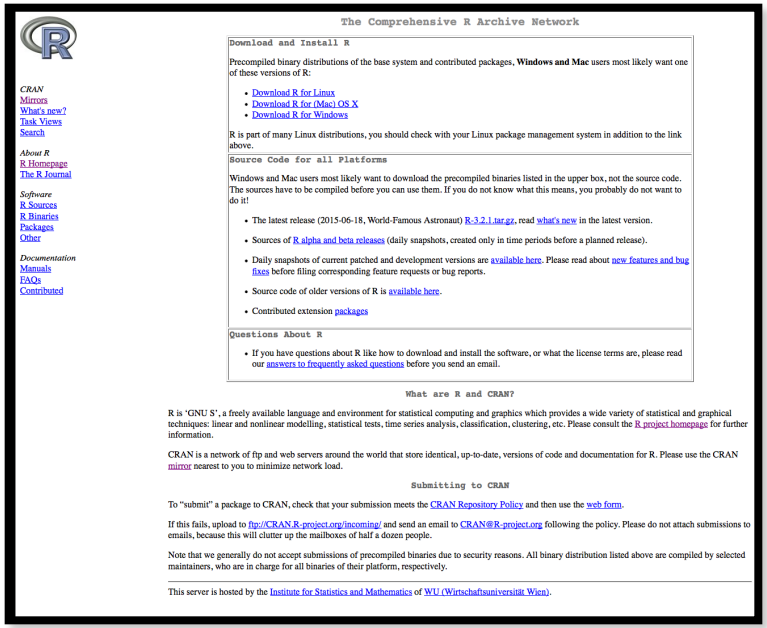
- **R version 3.2.2 (Fire Safety) prerelease versions** will appear starting 2015-08-04. Final release is scheduled for 2015-08-14.
- **The R Journal Volume 7/1** is available.
- **R version 3.2.1 (World-Famous Astronaut)** has been released on 2015-06-18.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.
- **useR! 2015**, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- **useR! 2014**, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.



CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

0-Cloud	
https://cran.rstudio.com/	Rstudio, automatic redirection to servers worldwide
http://cran.rstudio.com/	Rstudio, automatic redirection to servers worldwide
Algeria	
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
http://cran.csiro.au/	CSIRO
http://cran.ms.unimelb.edu.au/	University of Melbourne
Austria	
https://cran.r-project.org/	Wirtschaftsuniversität Wien
http://cran.at.r-project.org/	Wirtschaftsuniversität Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association
Brazil	
http://nbcgib.uesc.br/mirrors/cran/	Center for Comp. Biol. at Universidade Estadual de Santa Cruz
http://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
http://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
http://www.vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
http://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Canada	
http://cran.stat.sfu.ca/	Simon Fraser University, Burnaby
http://mirror.its.dal.ca/cran/	Dalhousie University, Halifax
http://cran.utstat.utoronto.ca/	University of Toronto
http://cran.skazkaforyou.com/	iWeb, Montreal
http://cran.parentingamerica.com/	iWeb, Montreal
Chile	
http://dirichlet.mat.puc.cl/	Pontificia Universidad Catolica de Chile, Santiago
China	
http://mirror.bjtu.edu.cn/cran/	Beijing Jiaotong University, Beijing
http://mirrors.opencas.cn/cran/	Chinese Academy of Sciences, Beijing
https://mirrors.tuna.tsinghua.edu.cn/CRAN/	TUNA Team, Tsinghua University
http://mirrors.tuna.tsinghua.edu.cn/CRAN/	TUNA Team, Tsinghua University
https://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
http://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
http://mirrors.xmu.edu.cn/CRAN/	Xiamen University
Colombia	
https://www.icesi.edu.co/CRAN/	Icesi University
http://www.icesi.edu.co/CRAN/	Icesi University



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2015-06-18, World-Famous Astronaut) R-3.2.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- 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.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Submitting to CRAN

To "submit" a package to CRAN, check that your submission meets the CRAN Repository Policy and then use the web form.

If this fails, upload to ftp://CRAN.R-project.org/incoming/ and send an email to CRAN@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a dozen people.

Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all binaries of their platform, respectively.

This server is hosted by the Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien).

1.4. Enlaces (2/2)

- R-Wiki :

  http://wiki.r-project.org/rwiki/doku.php

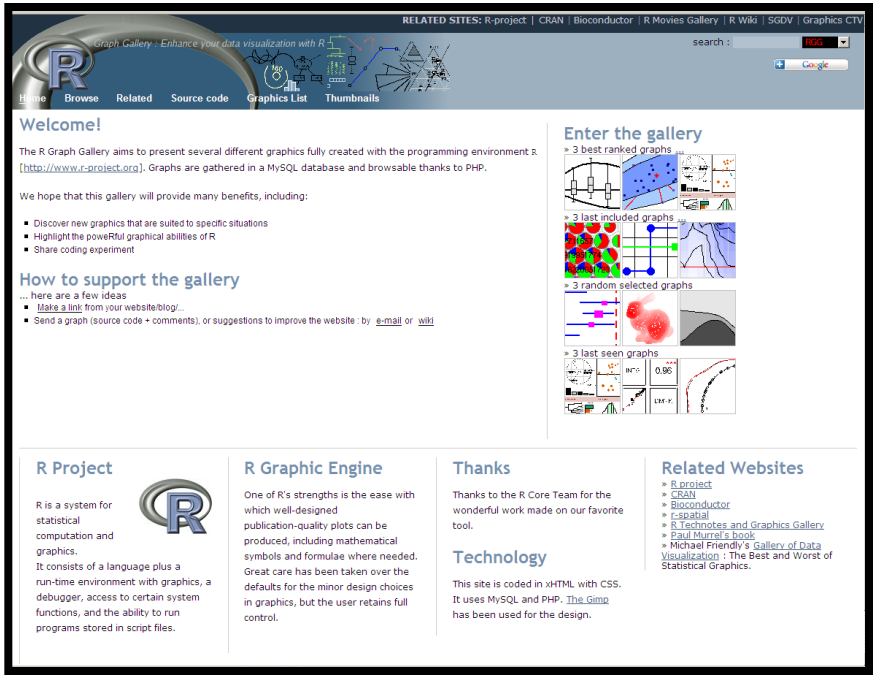
- Interfaz Web para R :

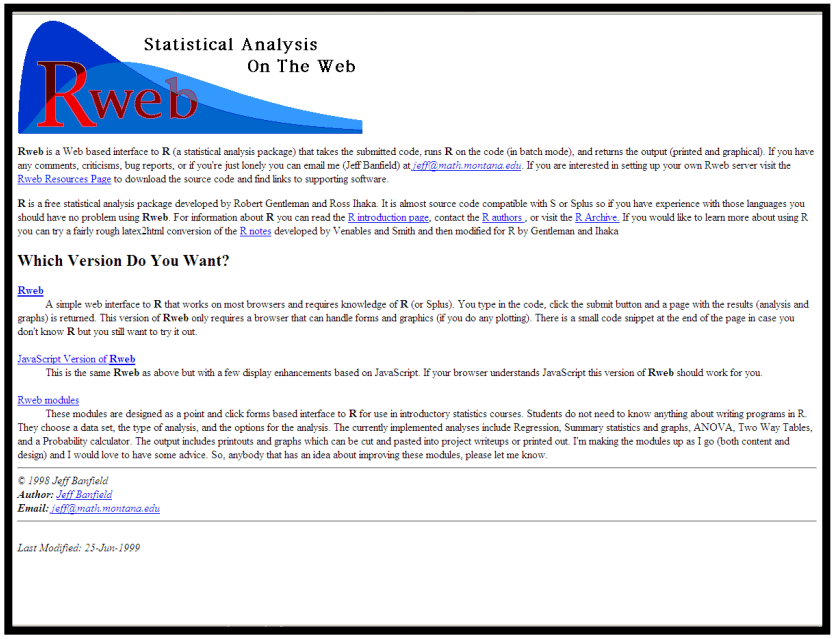
  http://www.math.montana.edu/Rweb/

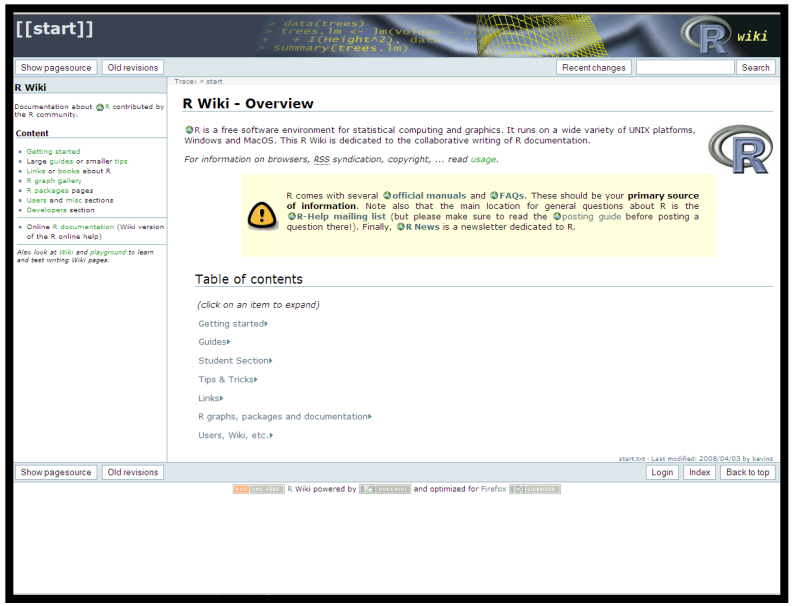
- R Graph Gallery :

  http://addictedtor.free.fr/graphiques/

Instituto de Salud Carlos III





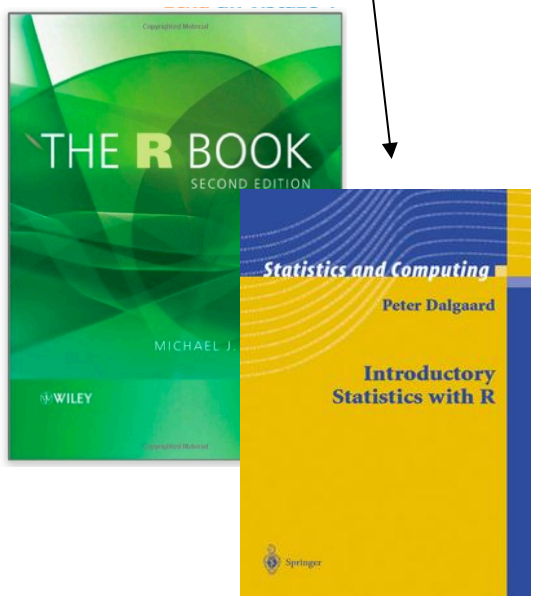


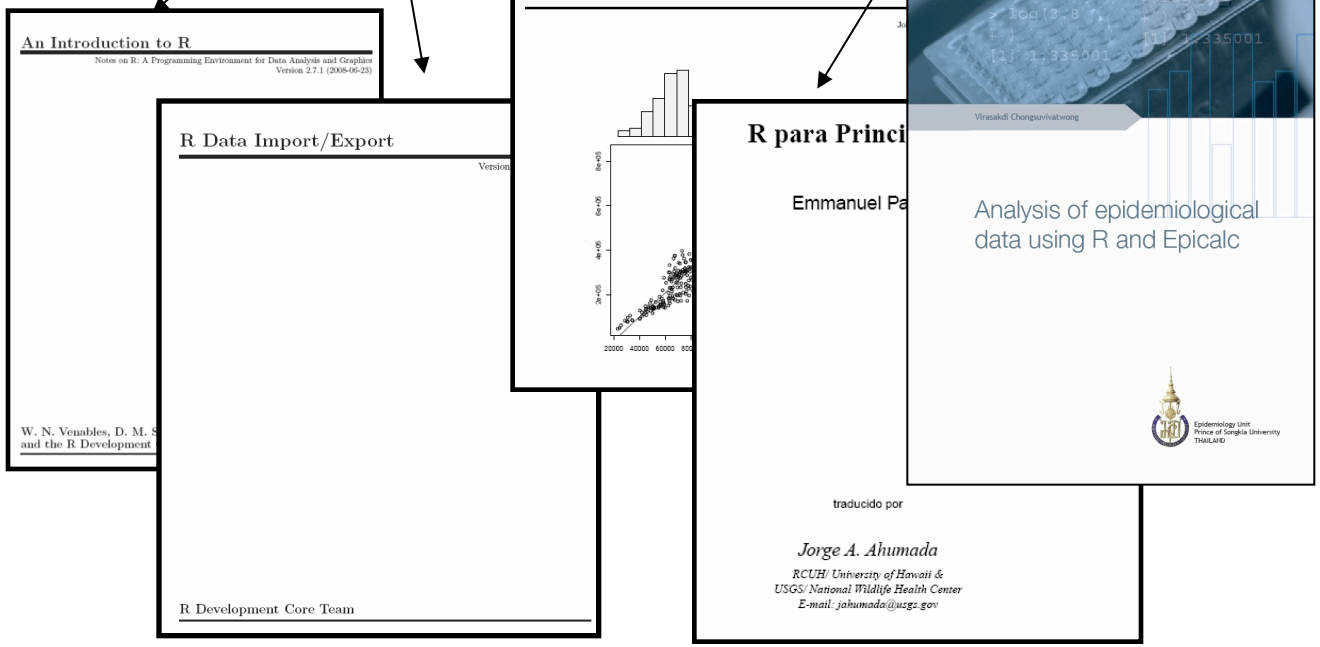
1.5. Libros y material general

Libros

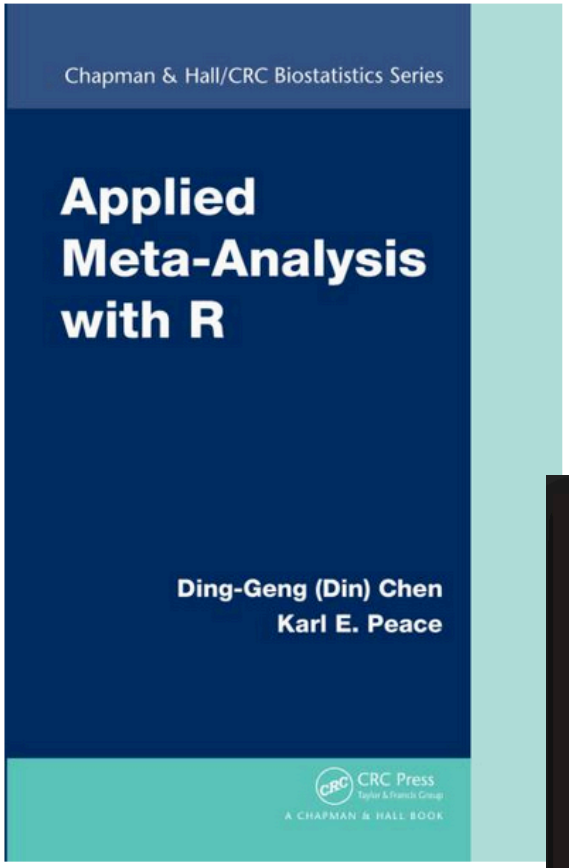
Manuales *R Core Team*

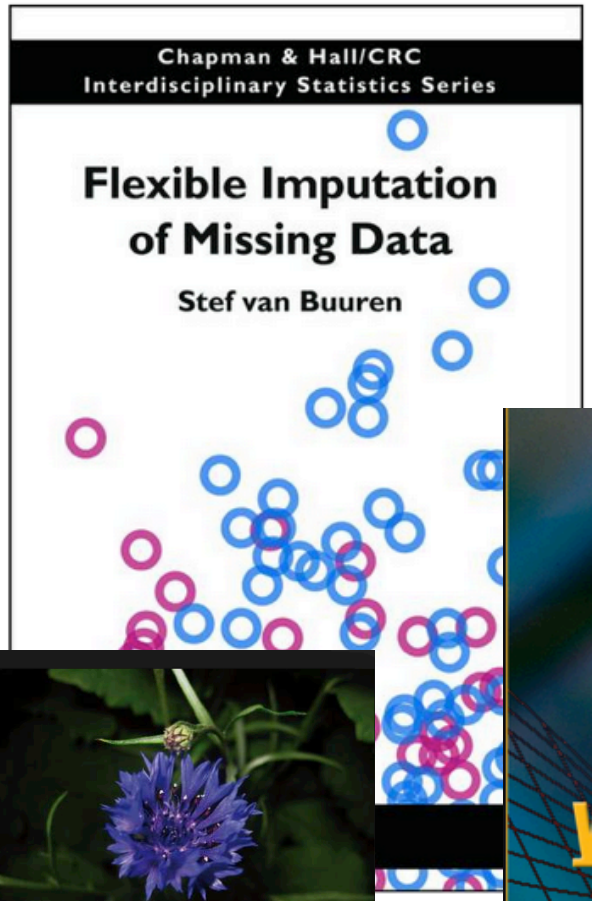
Documentación

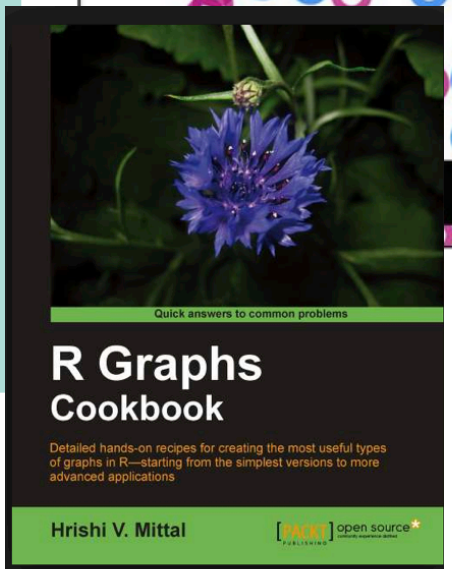


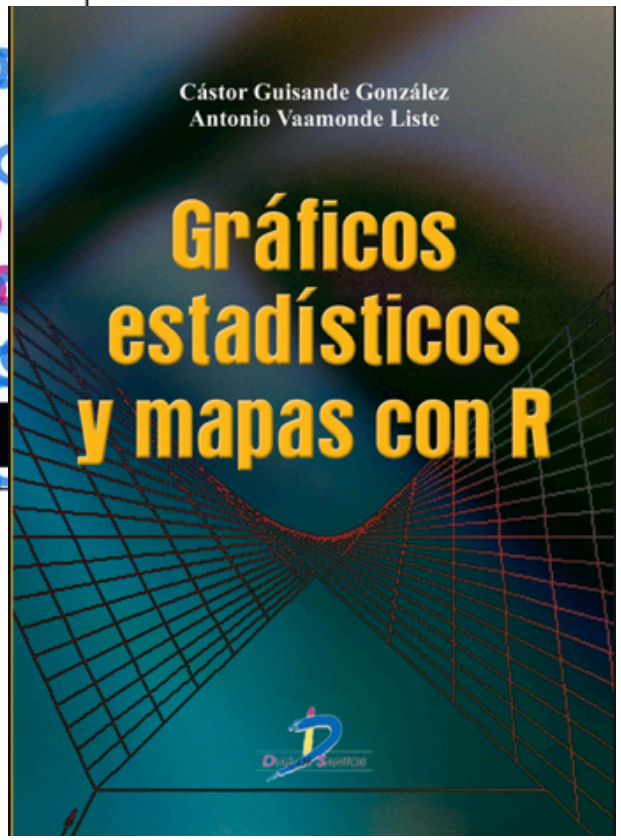


Instituto de Salud Carlos III

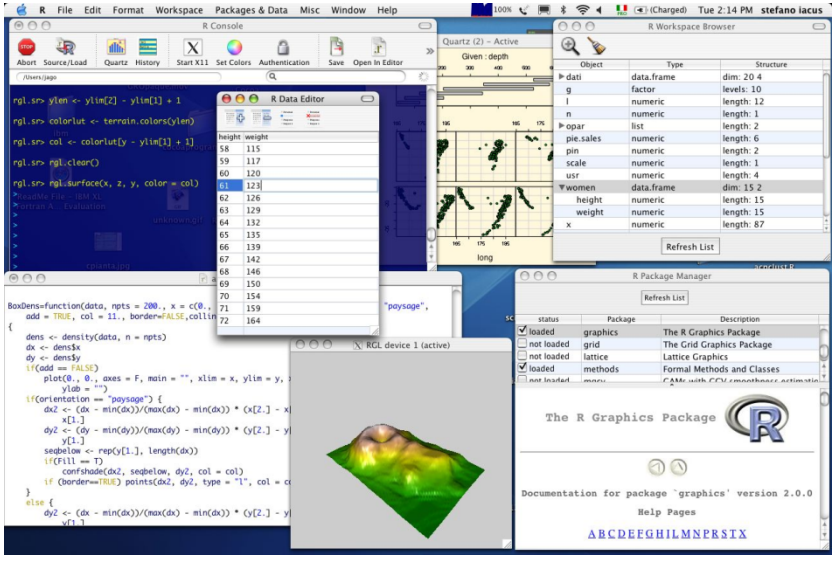


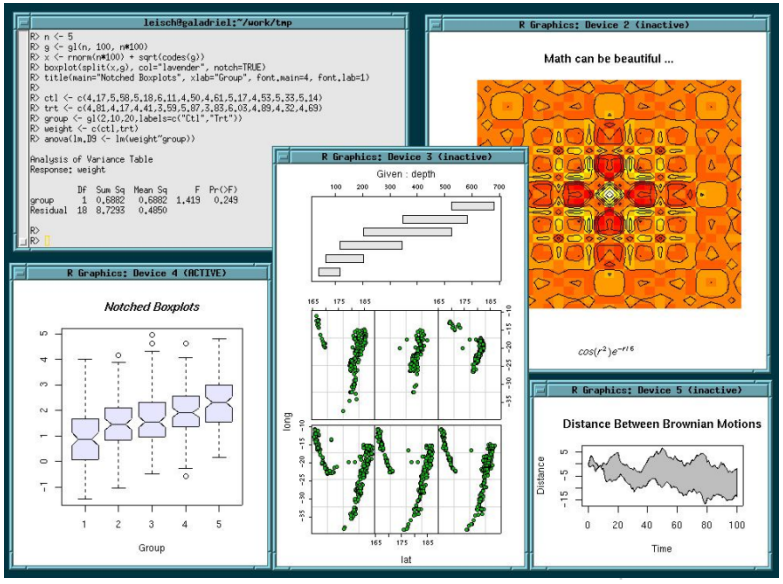


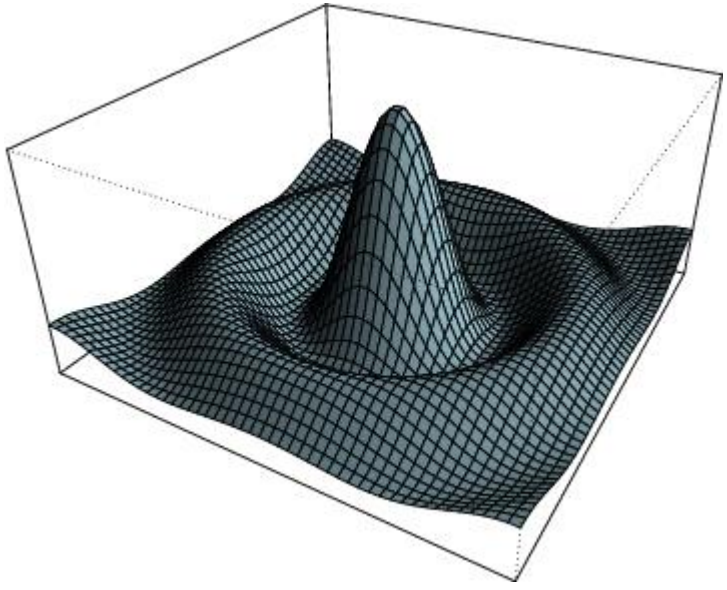


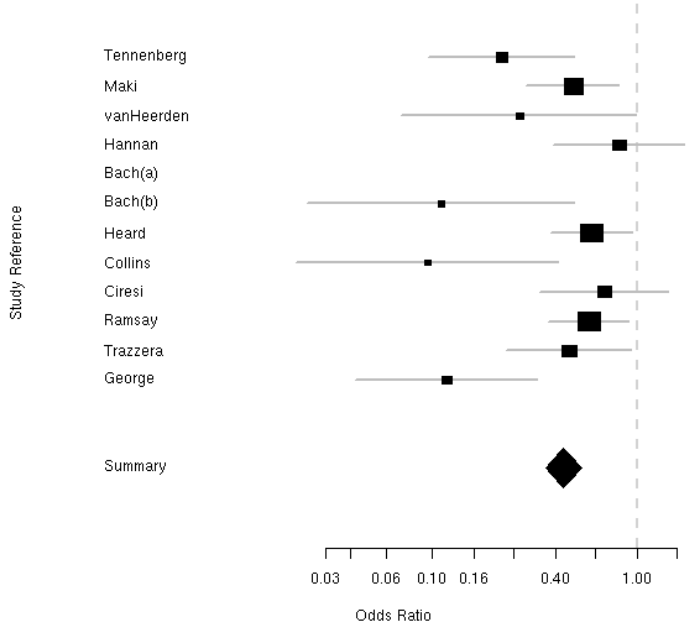


1.6. Screenshots









1.7. Comunidad R

- Burns Statistics Ltd., London, U.K.
- Department of Statistics, Brigham Young University, Utah, USA
- Paul von Eikeren, USA
- Institute of Mathematical Statistics (IMS), Ohio, USA
- Loyalty Matrix Inc., California, USA
- Mango Solutions, Chippenliam, UK
- Marc Schwartz, USA
- Merck and Co. Inc., USA
- Numbers Internation Pty Ltd, Australia
- Prediction Company, Santa Fe, New Mexico, USA
- Saxo Bank, Denmark
- Schröder Investment Management Ltd., London, UK
- InterContinental Hotels Group, USA
- Shell Statistics and Chemometrics, Chester, UK
- Statisticon AB, Uppsala, Sweden
- Astra Zeneca R&D Mölndal, Mölndal, Sweden
- AT&T Labs, New Jersey, USA
- Baxter AG, Vienna, Austria
- Baxter Healthcare Corp., California, USA
- BC Cancer Agency, Vancouver, Canada
- Black Mesa Capital, Santa Fe, USA
- Boehringer Ingelheim Austria GmbH, Vienna, Austria
- Breast Center at Baylor College of Medicine, Houston, Texas, USA
- Center für digitale Systeme, Freie Universität Berlin, Germany
- Dana-Farber Cancer Institute, Boston, USA
- Department of Biostatistics, Johns Hopkins University, Maryland, USA
- Department of Biostatistics, Vanderbilt University School of Medicine, USA
- Department of Economics, Stockholm University, Sweden
- Department of Mathematics and Statistics, Utah State University, USA
- Department of Statistics, University of California at Los Angeles, USA
- Department of Statistics, University of Warwick, Coventry, UK
- Department of Statistics, University of Wisconsin-Madison, Wisconsin, USA
- Department of Statistics, Iowa State University, USA
- Department of Statistics & Actuarial Science, University of Iowa, USA
- Dipartimento di Statistica, Università Cà Foscari di Venezia, Italy
- Division of Biostatistics, University of California, Berkeley, USA
- Ef-prime Inc., Tokyo, Japan
- European Bioinformatics Inst., UK
- Hygeia Associates, California, USA
- Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse, University of Augsburg, Germany
- MPI for Demographic Research, Rostock, Germany
- Massachusetts General Hospital Biostatistics Center, Boston, USA
- National Public Health Institute, Helsinki, Finland
- Norwegian Institute of Marine Research, Bergen, Norway
- School of Economics and Finance, Victoria University of Wellington, New Zealand
- Spotfire, Massachusetts, USA
- TERRA Lab, University of Regina - Department of Geography, Canada
- ViaLactia Biosciences (NZ) Ltd, Auckland, New Zealand
- Adelchi Azzalini (Italy)
- AT&T Research (USA)
- Austrian Association for Statistical Computing (Austria)
- BC Cancer Agency, Vancouver (Canada)
- Fabian Barth (Germany)
- Biostatistics and Research Decision Sciences, Merck Research Laboratories (USA)
- Brian Caffo (USA)
- David W. Crawford (USA)
- Dianne Cook (USA)
- Yves DeVille (France)
- Department of Economics, University of Milano (Italy)
- Dipartimento di Scienze Statistiche e Matematiche di Palermo (Italy)
- Emanuele De Rinaldis (Italy)
- Zubin Dowlaty (USA)
- Faculty of Economics, University of Groningen (Netherlands)
- Jaimison Fargo (USA)

1.8. Actualidad

- El proyecto R mantiene dos series de Conferencias organizadas regularmente por los miembros de la Comunidad de R:

  1) **useR!**== forum de la comunidad de usuarios de R.

  2) **DSC**== una plataforma de desarrollo de software estadístico.

Instituto de Salud Carlos III

International R User Conference

- Usuarios y desarrolladores de R.
- Presentaciones de invitados y usuarios contribuidores.
- Abstracts de usuarios
- Serie histórica:

useR! 2004, Vienna, Austria
useR! 2006, Vienna, Austria
useR! 2007, Ames, IA, USA
useR! 2008, Dortmund, Germany
useR! 2009, Rennes, France
useR! 2010, Gaithersburg, MD, USA
useR! 2011, Coventry, UK
useR! 2012, Nashville, TN, USA
useR! 2013, Albacete, Spain
useR! 2014, Los Angeles, CA, USA
useR! 2015, Aalborg, Denmark.
useR! 2016, Stanford, CA, USA
useR! 2017, Brussels, Belgium.
useR! 2018, Brisbane, Australia.
useR! 2019, Toulouse, France.
useR! 2020, Boston, MA, USA.

Instituto de Salud Carlos III

useR!

The R User Conference 2008

August 12-14, Technische Universität Dortmund, Germany

Organizer: Fakultät Statistik, Technische Universität Dortmund
Co-Organizer: Austrian Association for Statistical Computing
Sponsors: R Foundation for Statistical Computing

technische universität dortmund
mangosolutions data analysis that delivers
Lufthansa A STAR ALLIANCE MEMBER
MERCK & CO., INC. Whitehouse Station, N.J., U.S.A.
WILEY-BLACKWELL
CAMBRIDGE UNIVERSITY PRESS
statistics.com
ZEMENTIS ADAPA Predictive Analytics Edition
StatSoft
AASC Austrian Association for Statistical Computing
CRC Press Taylor & Francis Group CHAPMAN & HALL
REvolution computing
Technische Universität Dortmund SFB 475 Reduction of Complexity in Multivariate Data Structures
Google
Springer

Conference
About the Conference
Date & Location
Important Dates
Call for Papers
Download: Logo, Flyer, Poster
Funding
Participants
Photos

Program
Conference Program
Download (PDF): Program and Book of abstracts
Invited Lectures
Presentations (including slides!)
Tutorials
Social Program
Program Committee

Dortmund
Accommodation
About Dortmund
Travel information

About the Conference

useR! 2008, the R user conference, takes place at the Fakultät Statistik, Technische Universität Dortmund, Germany from 2008-08-12 to 2008-08-14. Pre-conference tutorials take place on August 11.
The conference is organized by the Fakultät Statistik, Technische Universität Dortmund and the Austrian Association for Statistical Computing (AASC). It is funded by the R Foundation for Statistical Computing.

Following the successful useR! 2004, useR! 2006, and useR! 2007 conferences, the conference is focused on

1. R as the `lingua franca' of data analysis and statistical computing,
2. providing a platform for R users to discuss and exchange ideas how R can be used to do statistical computations, data analysis, visualization and exciting applications in various fields,
3. giving an overview of the new features of the rapidly evolving R project.

As for the predecessor conference, the program consists of two parts:

1. invited lectures discussing new R developments and exciting applications of R,
2. user-contributed presentations reflecting the wide range of fields in which R is used to analyze data.

A major goal of the useR! conference is to bring users from various fields together and provide a platform for discussion and exchange of ideas: both in the formal framework of presentations as well as in the informal part of the conference in Dortmund's famous beer pubs and restaurants.

Prior to the conference, on 2008-08-11, there are tutorials offered at the conference site. Each tutorial has a length of 3 hours and takes place either in the morning or afternoon.

Organizing Team:

Uwe Ligges (conference), Achim Zeileis (program), Claus Weihs, Gerd Kopp (local organization), Friedrich Leisch, Torsten Hothorn

Program committee:

Date & Location
August 12-14, 2008 (iCalendar file)

Fakultät Statistik
Technische Universität Dortmund
Vogelpothsweg 87
44227 Dortmund
Germany

Important Dates
2008-03-31
early registration deadline
2008-03-31
submission deadline of abstracts
2008-05-18
notification of acceptance
2008-05-31(CEST)
regular registration deadline
closed now:
online registration
2008-07-25
registration deadline
(later registration possible on site)
2008-08-11
tutorials
2008-08-12
conference start
2008-08-14
conference end

Instituto de Salud Carlos III

useR!

Abstracts and slides sorted by author

Click [A] for abstracts or [S] for slides.

	Author	Title
[A]	Thomas Achia, Atinuke Adebanji, John Owino, Anne Wangombe	Spatial Durbin Model for Poverty Mapping and Analysis
[A][S]	Daniel Adler, Jens Oehlschlägel, Oleg Nenadic, Walter Zucchini	Large atomic data in R: package 'ff'
[A]	Claudio Agostinelli	Robust Inference in Generalized Linear Models
[A]	Arthur Allignol, Jan Beyersmann, Martin Schumacher	mvna, a R-package for the Multivariate Nelson-Aalen Estimator in Multistate Models
[A]	Micah Altman, Michael McDonald	BARD: Better Automated Redistricting
[A]	Gianmarco Altoè	The 'deltaR' package: a flexible way to compare regression models on independent samples using a bootstrap approach
[A]	M. Rui Alves, M. Beatriz Oliveira	Automatic construction of graphical outputs of common multivariate analyses with a special reference to predictive biplots
[A][S]	Thomas Baier	R in Automation: Accessing Real-time-data
[A]	Mehmet Balcilar	RSTAR: A Package for Smooth Transition Autoregressive Modeling Using R
[A][S]	Marco Ballin, Giulio Barcaroli	Tree-based and GA tools for optimal sampling design
[A][S]	Susana Barbosa	ArDec: Autoregressive-based time series decomposition in R
[A][S]	Benjamin Barnes, Karen Steindorf	Visualizing multivariate categorical and continuous data from epidemiologic studies: An expanded scatter plot matrix
[A][S]	Jan Beyersmann, Arthur Allignol, Martin Schumacher	Understanding product integration
[A][S]	Thomas Binsl, Jaap Heringa, David Alders, Hans van Beek	FluxEs: A Network

Segmented Poisson Models

Vidal E1,2, Pastor-Barriuso R1,2, **Pollan M1,2, Lopez-Abente G1,2**.

1 Environmental and Cancer Epidemiology Unit, National Centre for Epidemio
Health. Madrid, Spain.
2 CIBERESP, Spain.

Standard dose-response analyses (such as categorical, spline, or nonparar
flexible tools to describe the overall shape of the dose-response relation across
but the identification of trend changes with these methods is subjective. Speci
formally test for the existence of change-points in risk trends.

We propose a log-linear model for aggregated data with Poisson variance and
in which the predictor function consists of two intersecting straight lines connect
point through a hyperbolic transition function, that allows for abrupt changes o
between the linear trends. The model, that was implemented as an R function,
existence of a change-point, as well as point and interval estimates for its loca
and above it.

ESTIMATION OF STANDARD ERRORS IN NON-LINEAR REGRESSION MODELS: SPATIAL VARIATION IN RISK AROUND PUTATIVE SOURCES

Ramis, Rebeca1,2,3; Diggle, Peter1; **López-Abente, Gonzalo2,3**

1 Department of Medicine, Lancaster University, UK.
2 Cancer and Environmental Epidemiology Area, National Centre for Epidemiology. Carlos III Institute of Health, Madrid, Spain.
3 CIBERESP

Background
We consider the problem of investigating spatial variation in the risk of non-infectious diseases in populations exposed to pollution from one or more point sources.
The data most commonly available to study this question include case-counts (*Oi*) in each of a set of areas that partition the geographical region of interest, suitable denominators, Ei, proportional to the expected number of cases in each area, and the locations of the relevant point sources, from which we can compute distances *dij* between the *j*th focus and a reference location, typically the centroid, within the *I*th area. Also available in most applications are covariates relating to socio-economic status or other risk-factors associated with each area, which we denote by *Zk*.

The standard approach to the analysis of data of this kind is a log-linear regression of the case-counts on the covariates, with log-transformed denominators as an offset variable. To model distance-related point source effects, a log-linear formulation is unrealistic because of the need to combine an elevated risk close to the source with a neutral long-distance effect. We therefore extend the model by including a non-linear distance function, *f(dij)*,

Instituto de Salud Carlos III

useR!

The R User Conference 2013

July 10-12 2013

University of Castilla-La Mancha, Albacete, Spain

Supported by the R Foundation for Statistical Computing

Instituto de Salud Carlos III

Instituto de Salud Carlos III

- **Directions in Statistical Computing**
- Desarrolladores de software estadístico e investigadores en cómputo estadístico
- No centrado únicamente en R
- Serie histórica:

DSC 1999, Vienna, Austria

DSC 2001, Vienna, Austria

DSC 2003, Vienna, Austria

DSC 2005, Seattle, USA:

DSC 2007, Auckland, New Zealand

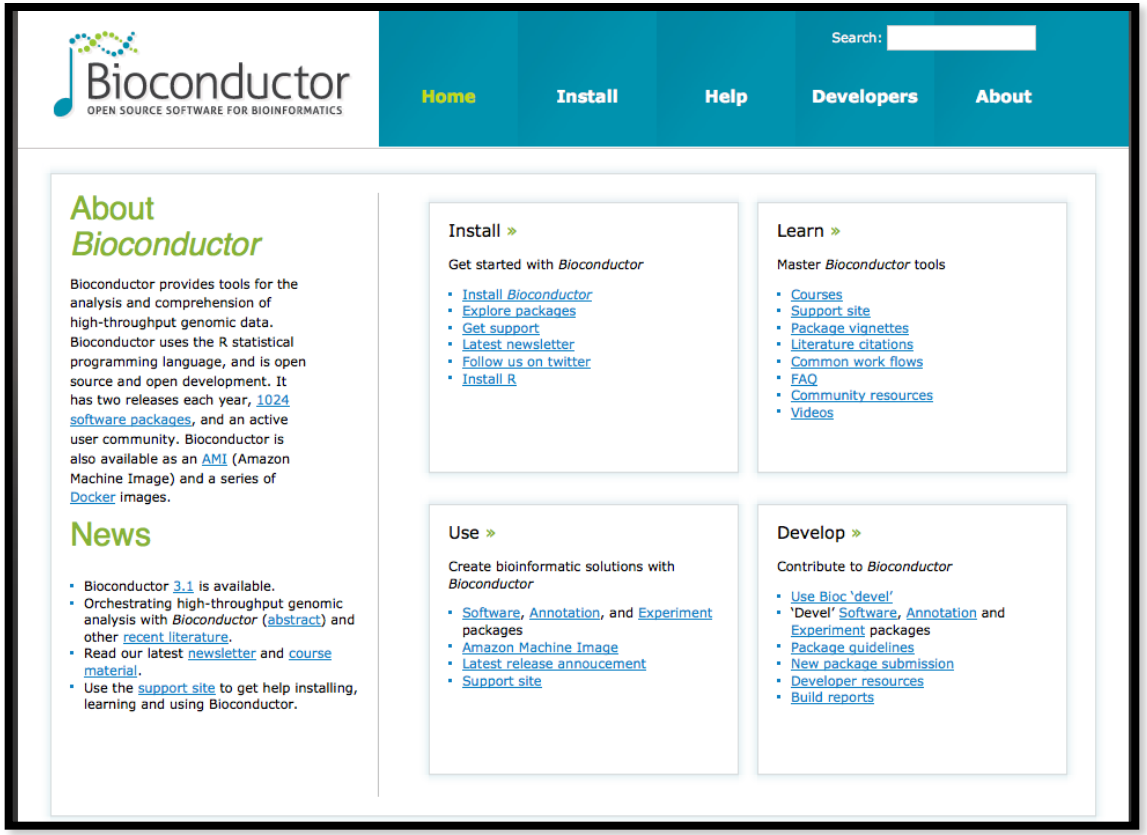
DSC 2009, Copenhagen, Denmark

Instituto de Salud Carlos III

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Open source, open development software project to provide tools for the analysis and comprehension of high-throughput genomic data.



http://www.bioconductor.org/

Instituto de Salud Carlos III

Package 'EBImage'

August 12, 2015

Version 4.10.1

Title Image processing and analysis toolbox for R

Encoding UTF-8

Author Andrzej Oleś, Gregoire Pau, Mike Smith, Oleg Sklyar, Wolfgang Huber, with contributions from Joseph Barry and Philip A. Marais

Maintainer Andrzej Oleś <andrzej.oles@embl.de>

Depends

Imports BiocGenerics (>= 0.7.1), methods, graphics, grDevices, stats, abind, tiff, jpeg, png, locfit, fftwtools (>= 0.9-7)

Suggests BiocStyle

Description EBImage provides general purpose functionality for image processing and analy-
nput) microscopy-based cellular assays, EBImage of-
ract quantitative cellular descriptors. This allows the automa-
gramming language and facili-
environment for signal processing, statistical modeling, ma-
on.

pbdR:

programming with big data in R

Simplifying Scalability

About | Packages | Download | Install | Tutorials | FAQ | Misc

Overview

The "Programming with Big Data in R" project (pbdR) enables high-level distributed data parallelism in R, so that it can easily utilize large HPC platforms with thousands of cores, making the R language scale to unparalleled heights. We interpret big data quite literally to mean that its size requires parallel processing either because it does not fit in the memory of a single multicore machine or because we need to make its processing time tolerable.

We achieve this, in part, by providing a simple interface to scalable, high performance libraries, such as MPI, ScaLAPACK, and NetCDF4. The routines in these libraries are engaged through R's classes and methods, so that the R language syntax is largely preserved, but with new, scalable, compiled code underneath. Most of the cumbersome distributed details are abstracted away for the user, although they are readily accessible should the user desire them.

We provide several R packages which focus on analyzing big data, especially on large distributed machines. The packages are meant to be used in the Single Program/Multiple Data (SPMD) programming model (see also), with special focus on large scale computing clusters. Packages and instructions for installation are available on the downloads page.

RDataMining.com: R and Data Mining

Search this site

Home
News
Training
Past Trainings
Documents
Introduction to Data Mining with R
R Reference Card for Data Mining
R and Data Mining: Examples and Case Studies
Introduction to Data Mining with R and Data Import/Export in R
Data Exploration and Visualization with R
Regression and Classification with R
Data Clustering with R
Association Rule Mining with R
Text Mining with R: Twitter Data Analysis
Time Series Analysis and Mining with R
Examples
Data Exploration
Decision Trees
Random Forest
k-means Clustering
Hierarchical Clustering
Outlier Detection
Time Series Forecasting
Time Series Analysis
Time Series Clustering and Classification
Association Rules
Text Mining
Twitter Follower Map
Social Network Analysis
Multidimensional Scaling (MDS)
Principal Component Analysis (PCA)
Parallel Computing
Other Examples
Big Data
Big Data Platforms

This website presents documents, examples, tutorials and resources on R and data mining.

Documents on Data Mining with R

- R Reference Card for Data Mining
- R and Data Mining: Examples and Case Studies
- Introduction to Data Mining with R
- RDataMining slides series on
  - Introduction to Data Mining with R and Data Import/Export in R
  - Data Exploration and Visualization with R,
  - Regression and Classification with R,
  - Data Clustering with R,
  - Association Rule Mining with R,
  - Text Mining with R: Twitter Data Analysis, and
  - Time Series Analysis and Mining with R

Examples on Data Mining with R

- data exploration
- decision trees
- k-means clustering and hierarchical clustering
- outlier detection
- time series decomposition and forecasting
- time series clustering and classification
- association rules
- text mining
- social network analysis
- Multidimensional Scaling (MDS)
- parallel computing
- many examples from other websites

News

www2.rdatamining.com: a mirror site for Chinese users

RDataMining.com now has a mirror website at http://www2.rdatamining.com. Users in China can download RDataMining documents, code and data at above mirror site, if no access to ...

Posted Aug 5, 2015, 5:31 AM by Yanchang Zhao

Showing posts 1 - 1 of 73. View more »

Sponsors

Call for Sponsors

Media Partners

BigDataSummit Data2Decision

Canberra Data Miners

AusDM 2015

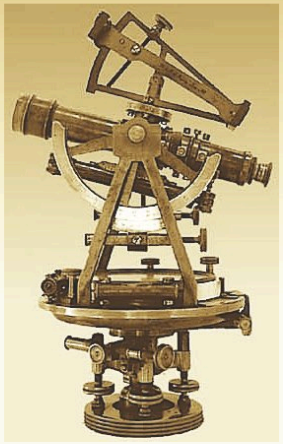
Sydney, 8-9 August 2015

BIGDATA WORLD SHOW 9th - 10th November 2015 Abu Dhabi, UAE

BIGIT TECHNOLOGY SHOW

Instituto de Salud Carlos III

Survey analysis in R



This is the homepage for the "survey" package, which provides facilities in R for analyzing data from complex surveys. The current version is 3.29. A much earlier version (2.2) was published in Journal of Statistical Software

An experimental package for very large surveys such as the American Community Survey can be found here

A port of a much older version of the survey package (version 3.6-8) to S-PLUS 8.0 is available from CSAN (thanks to Patrick Aboyoun at Insightful).

Features:

- Means, totals, ratios, quantiles, contingency tables, regression models, loglinear models, survival curves,rank tests, for the whole sample and for domains.
- Variances by Taylor linearization or by replicate weights (BRR, jackknife, bootstrap, multistage bootstrap, or user-supplied)
- Multistage sampling with or without replacement.
- PPS sampling with or without replacement: Horvitz-Thompson and Yates-Grundy estimators and a range of approximations.
- Post-stratification, generalized raking/calibration, GREG estimation, trimming of weights.
- Two-phase designs. Estimated weights for augmented IPW estimators.
- Graphics
- Support for using multiply imputed data
- Database-backed design objects for large data sets (now with replicate weights, too)
- Some support for parallel processing on multicore computers.
- Multivariate analysis: principal components, factor analysis (experimental).
- Likelihood ratio (Rao-Scott) tests for glms, Cox models, loglinear models.

The NEWS file gives a history of features and bug fixes.

Comparison shopping:
Alan Zaslavsky keeps a comprehensive list of survey analysis software for the ASA Section on Survey Research Methods.

User-generated ratings and reviews of this package (and others) at crantastic.

---

http://r-survey.r-forge.r-project.org/survey/

Instituto de Salud Carlos III

Methodology article

Open Access

Gene selection and classification of microarray data using random forest

Ramón Díaz-Uriarte*[1] and Sara Alvarez de Andrés[2]

Address: [1]Bioinformatics Unit, Biotechnology Programme, Spanish National Cancer Centre (CNIO), Melchor Fernandez Almagro 3, Madrid, 28029, Spain and [2]Cytogenetics Unit, Biotechnology Programme, Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Email: Ramón Díaz-Uriarte* - rdiaz@ligarto.org; Sara Alvarez de Andrés - salv

* Corresponding author

Environment International 40 (2012) 116–127

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Risk of dying of cancer in the vicinity of multiple pollutant sources associated with the metal industry

Javier García-Pérez [a,b,*], María Felicitas López-Cima [a,b,c], Marina Pollán [a,b], Beatriz Pérez-Gómez [a,b], Nuria Aragonés [a,b], Pablo Fernández-Navarro [a,b], Rebeca Ramis [a,b], Gonzalo López-Abente [a,b]

[a] Cancer and Environmental Epidemiology Unit, National Center for Epidemiology, Carlos III Institute of Health, Avda. Monforte de Lemos, 5, 28029 Madrid, Spain
[b] CIBER Epidemiología y Salud Pública (CIBERESP), Spain
[c] Molecular Epidemiology of Cancer Unit, University Institute of Oncology, University of Oviedo, C/Fernando Bongera, s/n, 33006 Oviedo, Spain

Software

SignS: a parallelized, open-source, freely available, web-b for gene selection and molecular signatures for survival a censored data

Ramon Diaz-Uriarte

Address: Statistical Computing Team, Structural Biology and Biocomputing Programme, Spanish National Cancer Center Fernández Almagro 3, Madrid, 28029, Spain

Email: Ramon Diaz-Uriarte - rdiaz02@gmail.com

Published: 21 January 2008

BMC Bioinformatics 2008, 9:30 doi:10.1186/1471-2105-9-30

Received:
Accepted:

This article is available from: http://www.biomedcentral.com/1471-2105/9/30

Research article

The presence of the pilus locus is a clonal property among pneumococcal invasive isolates

Sandra I Aguiar, Isa Serrano, Francisco R Pinto, José Melo-Cristino and

Molecular, Faculdade de Medicina Universidade de Lisboa, Lisboa, Portugal

aserrano@fm.ul.pt; Francisco R Pinto - fpinto@fm.ul.pt; José Melo-rez@fm.ul.pt

Received: 31 October 2007
Accepted: 28 February 2008

471-2180/8/41

Science of the Total Environment 435–436 (2012) 66–73

Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Proximity to mining industry and cancer mortality

Pablo Fernández-Navarro [a,b,*], Javier García-Pérez [a,b], Rebeca Ramis [a,b,c], Elena Boldo [a,b], Gonzalo López-Abente [a,b]

[a] Cancer and Environmental Epidemiology Unit, National Center for Epidemiology, Carlos III Institute of Health, Avda. Monforte de Lemos, 5, 28029 Madrid, Spain
[b] CIBER Epidemiología y Salud Pública (CIBERESP), Spain
[c] School of Health and Medicine, Lancaster University, UK