

Análisis de datos con R

Rubén Fernández Casal



Big Data en PyMEs

Diciembre 2016

- 1 Introducción
- 2 Acceso y manipulación de datos
- 3 Análisis exploratorio
- 4 Modelado de datos
- 5 Informes y aplicaciones

- 1 Introducción
 - Etapas del proceso
 - El entorno estadístico R
 - Compañías que usan R
- 2 Acceso y manipulación de datos
- 3 Análisis exploratorio
- 4 Modelado de datos
- 5 Informes y aplicaciones

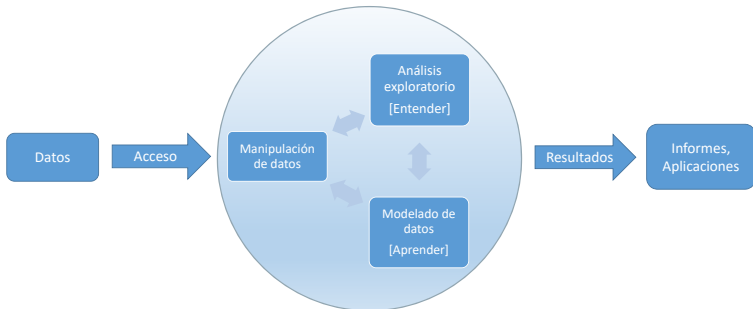
Introducción

Etapas del proceso

- **El objetivo es la obtención de información**

(para la toma de decisiones).

- El proceso:



- El entorno estadístico R puede ser una herramienta de gran utilidad en todo este proceso.

El entorno estadístico R

R es un lenguaje de programación desarrollado específicamente para el análisis estadístico y la visualización de datos.

- Lenguaje interpretado (similar a Matlab o Phytion)
 - derivado del S (Laboratorios Bell).
- Libre de código abierto (licencia GPL).
- Multiplataforma (Linux, Windows, MacOS, ...).

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



The R Project for Statistical Computing

[Home]

Download
CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting
Bugs
Development
Site
Conferences
Search

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

- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.

<http://cran.es.r-project.org>



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 (Monday 2016-10-31, Sincere Pumpkin Patch) [R 3.3.2 src.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha](#) and [beta releases](#) (daily snapshots).

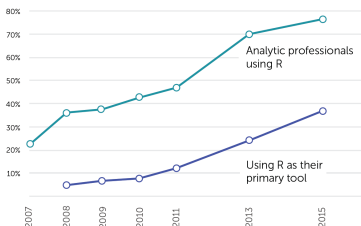
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQ
Contributed

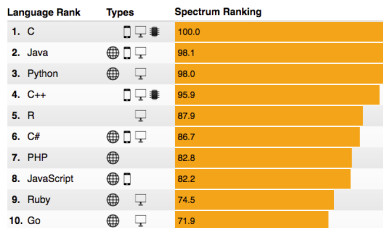
- Dispone de múltiples librerías (paquetes) que cubren “literalmente” todos los campos del análisis de datos.
 - Repositorios: CRAN (9705), Bioconductor (1289), ...
- Comunidad R muy dinámica (con muchas contribuciones).
 - R es muy popular...

Encuesta sobre el uso de R



Rexer Data Miner Survey 2007-2015

Popularidad lenguajes



IEEE Spectrum Top Programming Languages, 2016

- Puntos débiles (a priori): velocidad, memoria, ...

Ventajas de R respecto a otras alternativas

R destaca especialmente en:

- Representaciones gráficas.
- Métodos estadísticos “avanzados”

(*Data Science: Data Mining, Machine Learning, Statistical Learning, Business Intelligence, ...*):

- Datos funcionales.
 - Estadística espacial.
 - ...
- Análisis de datos “complejos”:
 - Big Data.
 - Lenguaje natural (*Text Mining*).
 - Análisis de redes.
 - ...

James *et al.* (2008). *An Introduction to Statistical Learning: with Applications in R*. Springer
(disponible en: <http://www-bcf.usc.edu/~gareth/ISL>).

Williams (2011). *Data Mining with Rattle and R*. Springer.

Compañías que usan R

Cada vez son más las empresas que utilizan R.

- **R Consortium**

Grupo de empresas que apoyan a la Fundación R y a la comunidad R.



- Otras compañías:

- Facebook, Twitter, Bank of America, Monsanto, ...
- <http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html>
- <http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html>



- Microsoft R Server

Diseñado para entornos Big Data y computación de altas prestaciones.

- Microsoft R Open

Versión de R con rendimiento mejorado.

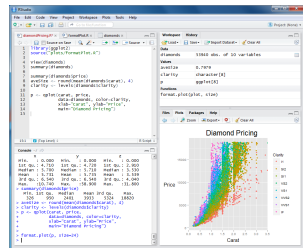
- Microsoft R Application Network:

<http://mran.revolutionanalytics.com>

- Integración de R con:

SQL Server, PowerBI, Azure y Cortana Analytics.

Microsoft no reveló el precio de compra;
la capitalización total de Revolution era de unos 40 millones de dólares en esa fecha.
<http://thomaswdinsmore.com/2015/01/26/microsoft-buys-revolution-analytics>



● RStudio Desktop

Entorno de desarrollo (IDE) con múltiples herramientas.

● RStudio Server

Interfaz web que permite ejecutar RStudio en el servidor.

- Evita el movimiento de datos a los clientes.
- Ediciones Open Source y Professional.

● Compañía muy activa en el desarrollo de R:

- Múltiples paquetes: Shiny, rmarkdown, knitr, ggplot2, dplyr, tidy, ...
- Hadley Wickham (Jefe científico de RStudio).

Van der Loo y de Jonge (2012). *Learning RStudio for R Statistical Computing*. Packt Publishing.

- 1 Introducción
- 2 Acceso y manipulación de datos
 - Ejemplo: Estadística Espacial
 - Ejemplo: datos en red
- 3 Análisis exploratorio
- 4 Modelado de datos
- 5 Informes y aplicaciones

Acceso a datos

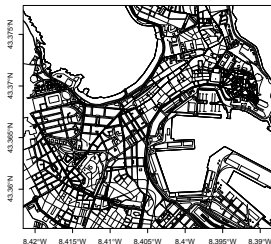
Hay una gran cantidad de paquetes de R para el acceso a distintos formatos/infraestructuras de datos:

- Archivos de datos
 - csv, xlsx, xml, json, sas, spss, stata, ...
- Bases de datos relacionales
 - DBI: Controladores de bases de datos nativos (máximo rendimiento).
 - RMySQL, ROracle, RSQLite, RPostgreSQL, ...
 - RODB, RJDBC, ...
- Big Data
 - Hadoop (RHadoop).
 - Hive (RHive), Spark (SparkR), ...
- Web scraping
 - rvest, RSelenium, twitterR, Rfacebook , ...

Munzert et al. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Wiley.

Ejemplo: Estadística Espacial

Por ejemplo, con R (`osmar`) se pueden importar mapas de OpenStreetMap y manipularlos fácilmente (`sp`, `igraph`, ...).



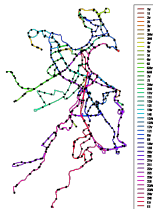
- Gran cantidad de paquetes para la manipulación y el análisis estadístico de datos espaciales (CRAN Task View: Analysis of Spatial Data):
 - El paquete `sp`
Bivand et al. (2008). *Applied Spatial Data Analysis with R*. Springer.
 - Integración con sistemas GIS, Maps y API de Google, ...

Manipulación de datos

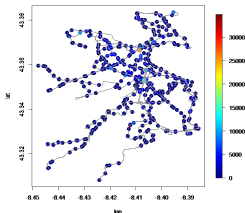
- El paquete básico de R proporciona múltiples herramientas:
 - `data.frame`: objeto de datos base (tabla de datos).
 - Para manipular conjuntos de datos (grandes) de forma más eficiente están disponibles otros paquetes:
 - `data.table`, `tibble`, `raster`, ...
 - El paquete `dplyr` proporciona un entorno homogéneo para el acceso y la manipulación de datos:
 - Utiliza una sintaxis de la forma:
`datos %>% filter %>% group_by %>% summarise...`
 - Puede trabajar con datos en distintos formatos:
 - `data.frame`, `data.table`, `tibble`, ...
 - Bases de datos relacionales
 - Hadoop (`plyrmr`), Spark (`sparklyr`), ...
- Evita emplear comandos de otros lenguajes (SQL, HQL, Scale, ...).

Ejemplo: datos en red

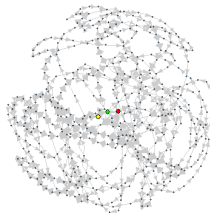
Red de autobuses de A Coruña



Los datos son atributos de nodos (paradas) y arcos (tramos)



Número de billetes



Nodos centrales

Kolaczyk y Csárdi (2014). *Statistical analysis of network data with R*. Springer.

- 1 Introducción
- 2 Acceso y manipulación de datos
- 3 **Análisis exploratorio**
 - Ejemplo: datos funcionales
- 4 Modelado de datos
- 5 Informes y aplicaciones

Análisis exploratorio

Métodos exploratorios (o de aprendizaje no supervisado):

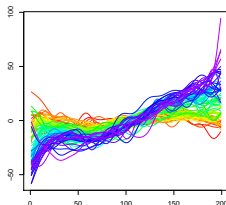
El objetivo es obtener información directamente de los datos, buscar relaciones y patrones, ...

(Tukey, 1975)

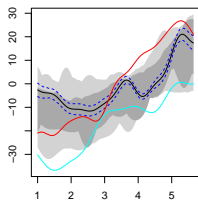
- Análisis descriptivo.
 - Gráficos:
 - Estáticos: estándar, lattice, ggplot2, ...
 - Dinámicos: rggobi, ggvis, ...
- Métodos de reducción de la dimensión:
 - Análisis de componentes principales, análisis factorial, ...
 - Análisis Clúster.
- Detección de datos atípicos.
- ...

Ejemplo: datos funcionales

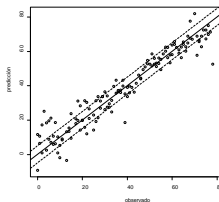
Mediciones durante la producción de silicio (parámetros eléctricos del horno).



Cada observación es una curva (correspondiente a una colada)



Boxplot funcional



Predicciones duración crisol (modelo funcional)

Ramsay *et al.* (2009). *Functional Data Analysis with R and MATLAB*. Springer.

- 1 Introducción
- 2 Acceso y manipulación de datos
- 3 Análisis exploratorio
- 4 Modelado de datos
 - Ejemplo: Modelos aditivos
- 5 Informes y aplicaciones

Modelado de datos

- La realidad puede ser muy compleja por lo que es habitual emplear un modelo para tratar de explicarla.
 - Modelos estocásticos (con componente aleatoria).
 - Tienen en cuenta la incertidumbre debida a no disponer de información suficiente.
 - La inferencia estadística proporciona herramientas para ajustar y contrastar la validez del modelo.
 - El objetivo es disponer de una aproximación simple de la realidad que sea útil (George Box: "En esencia, todos los modelos son falsos, pero algunos son útiles").
 - En ocasiones el objetivo es únicamente predecir.
- Métodos (de aprendizaje supervisado):
 - Clasificación
 - Análisis discriminante, Regresión logística, ...
 - Árboles de decisión, *bagging*, *random forest*, *boosting*
 - *Support vector machines* (SVM)
 - Regresión

Métodos de regresión

Algunas de las funciones y paquetes disponibles:

- Modelos paramétricos

- Modelos lineales:
 - Regresión lineal: `lm()` (`aov()`, `lme()`, `biglm`, ...).
 - Regresión lineal robusta: `rlm()` (`MASS`, ...).
 - Métodos de regularización (Ridge regression, Lasso): `glmnet`, ...
- Modelos lineales generalizados: `glm()` (`bigglm`, ...).
- Modelos paramétricos no lineales: `nls()` (`nlme`, ...).

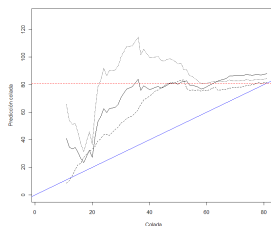
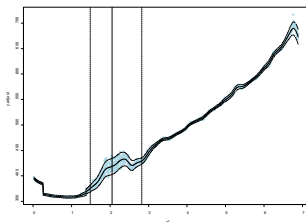
- Modelos no paramétricos

- Regresión local (métodos de suavizado):
`loess()`, `KernSmooth`, `sm`, ...
- **Modelos aditivos generalizados (GAM):** `gam`, `mgcv`, ...
- Árboles de decisión (Random Forest, Boosting):
`rpart`, `randomForest`, `xgboost`, ...
- Redes neuronales (`nnet`, ...), ...

Con todos los modelos se trabaja de una forma muy similar en R.

Ejemplo: Modelos aditivos

Mediciones durante la producción de silicio (parámetros eléctricos del horno).



Estimación de la media (y de la variabilidad) en una colada y predicciones del número total de coladas (derivadas del modelado del proceso).

- Modelo:

$$Y = \beta_0 + f_1(\mathbf{X}_1) + f_2(\mathbf{X}_2) + \cdots + f_p(\mathbf{X}_p) + \varepsilon,$$

con f_i , $i = 1, \dots, p$, funciones cualesquiera (suaves).

- Adicionalmente se puede considerar una función link.

Wood (2006). *Generalized Additive Models: An Introduction with R*. Chapman.

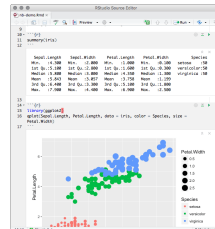
Indice

- 1 Introducción
- 2 Acceso y manipulación de datos
- 3 Análisis exploratorio
- 4 Modelado de datos
- 5 Informes y aplicaciones

Generación de informes



This screenshot shows the RStudio interface with two panes. The left pane displays R code chunks, including a chunk titled 'R Code Chunks' with code for plotting a scatter plot and a summary of a dataset. The right pane shows the rendered HTML output of this code, which includes a title 'R Code Chunks', a brief explanation of R Markdown, and a scatter plot of 'speed' vs 'dist' with a fitted curve. The plot is titled 'split(speed, dist, data = cars) + geom_smooth()'.



● Informes

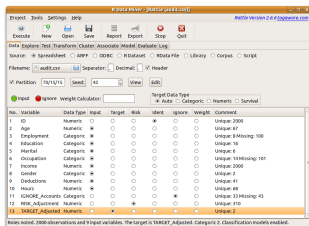
- R-Markdown permite la creación de informes en distintos formatos (HTML, PDF, DOCX, ...) de forma muy sencilla (rmarkdown).
- Otros paquetes: knitr, htmltools, ...

● Notebooks

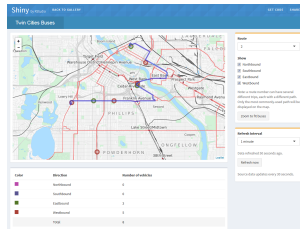
- Permiten la ejecución interactiva (generación automática de resultados).
- RStudio notebooks, Apache Zeppelin, Jupyter, ...

Yihui Xie (2015). *Dynamic Documents with R and knitr*. Chapman.

Desarrollo de aplicaciones



Rattle (Gtk)



Shiny

- Hay paquetes que permiten emplear librerías gráficas para la creación de aplicaciones:

- tcltk2, RGtk2, rJava, ...

Lawrence y Verzani (2012). *Programming Graphical User Interfaces in R*. Chapman and Hall/CRC.

- Shiny permite crear aplicaciones web interactivas fácilmente (incluso se pueden crear con rmarkdown).

- Otras alternativas: opencpu, httpuv, ...

Beeley (2015). *Web Application Development with R Using Shiny*. Packt Publishing.

Eso es todo...

