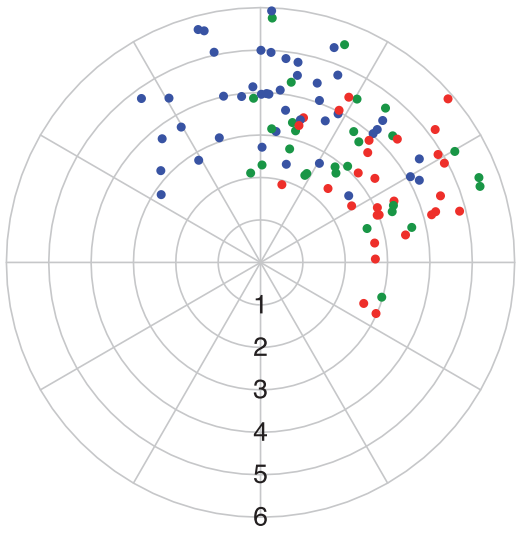
IV XORNADA DE USUARIOS DE R EN GALICIA



19 DE OUTUBRO DE 2017

SANTIAGO DE COMPOSTELA

www.r-users.gal

> LUGAR

Facultade de Matemáticas

> ORGANIZA

CIXUG

> COLABORA

USC Universidade de Santiago de Compostela — Facultade de Matemáticas

> PATROCINAN

amtega — Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

IV XORNADA DE USUARIOS R EN GALICIA

PROGRAMA
E RESUMOS

Santiago de Compostela

19 de Outubro de 2017

ORGANIZA:

Oficina de Software Libre (OSL) do CIXUG

Editora: Mª José Ginzo Villamayor

Outubro 2017

PRESENTACIÓN

A Oficina de Software Libre (OSL) do CIXUG comprácese en presentar a IV Xornada de Usuarios de R en Galicia.

Pretende ser un punto de encontro para todas aquelas persoas interesadas en intercambiar as súas experiencias e atopar colaboracións co resto da comunidade, ademais trátase de promocionar e difundir o coñecemento libre da linguaxe estatística R e mostrar as súas aplicacións.

O programa contempla doce relatorios ao longo de todo o día, ademais de dous obradoiros: Introdución a R e Web Scraping.

Entre os participantes figuran especialistas da Oficina de Software Libre do CIXUG, do Instituto Galego de Estatística, da Consellaría de Sanidade e outros organismos da Xunta de Galicia, das tres universidades galegas, do Cesga e de empresas como INDRA ou SolidQ.

Todo isto non sería posible sen o patrocinio de AMTEGA e a colaboración de da Facultade de Matemáticas, ás que agradecemos a súa contribución. Confiamos que os asistentes a xornada desfruten da mesma e dunha cidade que os acolle cos brazos abertos.

Santiago de Compostela, outubro de 2017

O Comité Organizador

COMITÉ ORGANIZADOR

Roberto Martín Souto
Oficina de Software Libre (CIXUG)

Rafael Rodríguez Gayoso
Concello de Santiago de Compostela

Mª José Ginzo Villamayor
Universidade de Santiago de Compostela

COMITÉ CIENTÍFICO

Mª José Ginzo Villamayor
Universidade de Santiago de Compostela

Miguel Ángel Rodríguez Muíños
Dirección Xeral de Saúde Pública (Consellería de Sanidade)

INFORMACIÓN XERAL

SEDE

Facultade de Matemáticas

Universidade de Santiago de Compostela

C/ Lope Gómez de Marzoa s/n

15782, Santiago de Compostela

DATAS

19 de outubro de 2017

ACCESO WIFI NA SEDE

SSID: usuariosr
Login: usuariosr
Password: T4rqbzpx2?

CERTIFICADOS

Todos os certificados se enviarán en formato dixital por correo electrónico unha vez rematada a Xornada.

UBICACIÓNS NA FACULTADE

Relatorios: Aula Magna

Talleres: Aulas 9 e 10 (nivel 4).

Cafés: corredor nivel 3.

PROGRAMA

09:45 - 10:10	Utilización do software R na unión de rexistros administrativos *Noa Veiguela Fernández (IGE)*
10:10 - 10:35	Cálculo da adxudicación dos premios de fin de carreira(2017) *Marcos Fernández Arias, Pablo Espido Noya (Xunta de Galicia)*
10:35 - 11:00	Xeración de música determinista con R *Miguel Ángel Rodríguez Muiños (Consellería de Sanidade); Alejandro Rodríguez Antolín*
11:30 - 11:55	Novo entorno para Big Data con R: sparklyr *Aurora Baluja González (CHUS); Javier Lopez Cacheiro (CESGA)*
11:55 - 12:20	R en paralelo e HPC Diego Mairena Díaz, Andrés Gómez, *Aurelio Rodríguez (CESGA); Santiago Cerviño (Instituto Español de Oceanografía)*
12:20 - 12:45	R como pedra angular de proxectos de Data Science *Daniel Prieto Rodríguez (Minsait by Indra)*
12:45 - 13:10	Predición de Series Temporais en Datasets Multidimensionais de Situacións de Negocio Mediante Paralelización Agradable en R Server *Antonio Soto Rodríguez (SolidQ)*
13:10 - 13:35	Análise da incidencia da leucemia granulocítica empregando a estimación non paramétrica de conxuntos de nivel *Paula Saavedra Nieves (UVigo)*
13:35 - 14:00	RStudio como ferramenta para desenvolvemento de material docente interactivo y multimedia *Alejandro Quintela del Río (UDC)*
16:00 - 16:25	Debuxando curvas ROC en R *Arís Fanjul Hevia (USC)*
16:25 - 16:50	A ecoloxía na súa revolución da cantidade masiva de datos *Miguel Branco (UVigo)*
16:50 - 17:15	Web Scraping *José Luis Juncal Pérez*
18:00- 20:00	Obradoiro: Introdución a R *Arís Fanjul Hevia, Mª José Ginzo Villamayor (USC)*
	Obradoiro: Web Scraping *José Luis Juncal Pérez*

Índice

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

UTILIZACIÓN DO SOFTWARE R NA UNIÓN DE REXISTROS ADMINISTRATIVOS

Noa Veiguela Fernández[1]

[1]Instituto Galego de Estatística (IGE)

RESUMO

Os rexistros administrativos constitúen unha fonte vital de información para os institutos de estatística pública. O aproveitamento destas bases de datos reduce a necesidade de recorrer a outras fontes, coma os censos e as enquisas; este feito trae aparellado a diminución, por unha banda, da carga sobre a poboación enquisada e, por outra, os custos que a elaboración destas operacións supoñen para os institutos de estatística. Ademais, se os datos polos que se interroga aos cidadáns xa están dispoñibles en rexistros administrativos, que sentido ten tratar de volver a obtelos por medio dun censo ou dunha enquisa? Non obstante, non todo son vantaxes á hora de traballar con rexistros administrativos. Hai que ter presente que a finalidade para a que se crean non é a labor estatística, polo que os conceptos e definicións incluídos neles rara vez coincidirán cos das estatísticas oficiais. Á hora de explotalos estatísticamente haberá que ter en conta tamén a normativa que está detrás da creación e xestión dos mesmos, e os posibles cambios que afecten a dita normativa. No Instituto Galego de Estatística embarcámonos no 2017 na ardua tarefa de casar todos os rexistros administrativos dos que dispomos, coa finalidade de crear unha macro base de datos que conteña información socioeconómica da poboación de Galicia. Esta base de datos contará, entre outros, cos seguintes datos:

- Variables que permitan determinar as características demográficas da persoa (nome e apelidos, sexo, idade, nacionalidade)
- A súa relación coa actividade económica: se traballou no ano de referencia da base de datos ou non, e as características dos distintos traballos que desempeñou (por conta propia ou allea e, neste caso, identificación da empresa, da modalidade contractual, o coeficiente de temporalidade, o sector de actividade, etc.)
- No caso de percibir algunha pensión contributiva da Seguridade Social no ano de referencia, as características da mesma
- Se a persoa se viu afectada por algún dos seguintes fenómenos demográficos no ano de referencia da base de datos: nacemento, defunción e/ou matrimonio
- Identificación das relacións de parentesco e de convivencia entre persoas (pais, nais, fillos, avós, avoas e persoas que habitan nun mesmo enderezo)
- Información xeoreferenciada do inmoble de residencia da persoa, así coma do lugar onde traballa

O rexistro de partida é o Padrón Municipal de habitantes do Instituto Nacional de Estatística, que contén información demográfica sobre toda a poboación que reside nalgún concello galego a unha data determinada. Para cubrir as restantes liñas informativas, contamos coa base de datos de afiliacións e contas de cotización á Seguridade Social, así como co rexistro de pensións do nomeado organismo, que nos subministra o Ministerio de Empleo y Seguridad Social. Tamén contamos, no caso dos traballadores non afiliados ao Sistema da Seguridade Social, cos rexistros de afiliación a distintas mutualidades á marxe do dito sistema. No referido á determinación dos fenómenos demográficos

que afectan a cada persoa, servímonos do Movemento Natural de Poboación do Instituto Nacional de Estadística (INE). E, finalmente, para xeoreferenciar cada inmoble, ben se trate da residencia dunha persoa ou do seu emprazamento de traballo, utilizamos indistintamente o Modelo de Direcciones de la Administración General del Estado (MDAGE) do INE, Cartociudad do Instituto Geográfico Nacional (IGN) e a información subministrada pola Dirección General del Catastro (entre outras). A priori é lóxico pensar que a variable de cruce por antonomasia no caso dos rexistros administrativos é o DNI, xa que se trata dun documento que permite identificar "unívocamente" a cada persoa. Pero, como se explicará ao longo da ponencia, non sempre se pode recorrer a esta variable no procedemento de unión por dous motivos:

- Nalgunhas bases de datos e rexistros administrativos non se dispón desta variable, ben porque o usuario non está obrigado a subministrala cando se inscribe no rexistro, ou ben porque o organismo xestor non nola ofrece por termos de protección de datos
- Constatamos que o DNI non permite identificar unívocamente a cada persoa, xa que, aínda que en raras ocasións, hai persoas que teñen asignado o mesmo número e tamén hai persoas que non dispoñen deste identificador

A finalidade desta ponencia consiste, precisamente, en presentar o procedemento do que nos servimos para lograr a unión dos rexistros administrativos cando non podemos contar co DNI da persoa. Nestes casos empregamos variables comúns en dous rexistros, coma o nome e apelidos da persoa, o seu sexo, data de nacemento e concello de residencia, que soen figurar en case todas as bases de datos que se manexan no IGE. Agora ben, para unha mesma persoa estas variables poden presentar discrepancias dun rexistro a outro; por exemplo, que nun rexistro se inclúa o segundo nome da persoa e no outro non ou que, ao gravar a data de nacemento, o xestor cometera unha errata e cambiase un dos díxitos da mesma. Para tratar de casar variables que poden diferir para unha mesma persoa dun rexistro a outro, empregáronse as librarías *stringdist* e *fuzzyjoin* do software libre de programación R. Ambas permiten comparar o grao de similitude que existe entre cadeas de texto ou entre variables numéricas dun xeito rápido e sinxelo, vinculando cada rexistro dunha base co rexistro da outra base de datos co que garde maior grao de semellanza.
Aproveitaremos tamén este foro para presentar os procedementos que se empregan no IGE para xeoreferenciar os lugares de residencia e de traballo das persoas. Trátase dunha liña de actuación, a da xeoreferenciación da información estatística, moi "en voga" nos últimos anos. Coñecer o emprazamento concreto de residencia da persoa, así coma o do lugar de traballo, permitirá analizar relacións entre localizacións xeográficas concretas, áreas ou zonas de influencia, que vaian máis alá das delimitacións que se soen empregar na estatística pública (concellos, ou, baixando un nivel, entidades de poboación). Esta información tamén pode axudar aos poderes públicos a trazar políticas de actuación focalizadas en territorios concretos. Neste procedemento de xeoreferenciación botamos man tamén do software libre R e dos diversos paquetes cartográficos que ofrece.

Palabras e frases chave: rexistros administrativos, stringdist, fuzzyjoin, xeoreferenciación.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

CÁLCULO DA ADXUDICACIÓN DOS PREMIOS DE FIN DE CARREIRA (2017)

Marcos Fernández Arias[1] e Pablo Espido Noya[1]

[1]Xunta de Galicia

RESUMO

Explicamos brevemente como se programa e calcula mediante **R** a adxudicación dos premios fin de carreira universitarios da Comunidade Autónoma de Galicia.

Palabras e frases chave: adxudicación de subvencións, premios fin de carreira, Galicia, seudonimización de datos persoais.

1. ANONIMIZACIÓN DE DATOS PERSOAIS

Para ponencias, tutoriais, exemplos e outras cesións de datos, resulta convinte en ocasións "seudonimizar"[1] previamente os datos persoais. De tal maneira que xa non se poden atribuír a un interesado se non se dispón de información adicional.
Explicamos brevemente un exemplo e métodos.

2. CÁLCULO DA ADXUDICACIÓN

Os criterios establecidos na actual convocatoria[2] de premios de fin de carreira, son diferentes aos dos anos anteriores. Por tal motivo realizouse un cálculo a medida.
Explicamos brevemente o procedemento utilizado segundo o cal se agrupan as titulacións en áreas e logo se elixen os candidatos de maior puntuación.

3. CONCLUSIÓNS

Mediante a utilización da linguaxe de programación **R** conséguese un aforro de tempo e unha maior fiabilidade nos resultados cando se precisan realizar cálculos numéricos diversos.
Neste caso a tarefa mostrada resultaba simple e asumible.
É un exemplo introductorio dos cálculos que se poden realizar mediante a linguaxe **R**.

Referencias

[1] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679

[2] Orde do 19 de xuño de 2017 pola que se establecen as bases reguladoras e se procede á convocatoria dos premios fin de carreira da Comunidade Autónoma de Galicia para o alumnado que rematou os seus estudos universitarios no ano 2016 nas universidades do Sistema universitario de Galicia (DOG nº 125 do 3 de xullo de 2017) http://www.xunta.gal/dog/Publicados/2017/20170703/AnuncioG0164-220617-0002_gl.html

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

GENERACIÓN DE MÚSICA DETERMINISTA CON R

Miguel Ángel Rodríguez Muíños[1] e Alejandro Rodríguez Antolín[2]

[1]Xunta de Galicia
[2]Universidad Autónoma de Madrid

RESUMO

A qué suena un cuadro?

Con motivo de la invitación a participar en los Encuentros Internacionales de Música Electroacústica "Monaco Electroacustique 2017" [1] celebradas en la Academia Rainiero III de Mónaco del 4 al 6 de Mayo de 2017, nos hemos planteado el reto de intentar presentar un proyecto experimental mediante el cual podamos "leer" (extraer información numérica de) un cuadro famoso (realmente cualquier imagen digitalizada) y ser capaces de transformarlo en música, mediante un proceso determinista (cada imagen genera su propia composición y siempre genera la misma -aunque se puede elegir el "estilo" de música a generar-). Este proceso ha sido desarrollado integramente en R y el resultado es un programa capaz de realizar lo anteriormente descrito. Este trabajo, que se presenta en las IV Xornadas de Usuarios de R en Galicia, consiste en la explicación del proyecto DETMUS [2] y en el comentario de aspectos técnicos de programación con R y sus utilidades fuera de los "usos tradicionales" del mencionado software.

Palabras e frases clave: música, determinista, R, gWidgets, tuneR, imageR.

Referencias

[1] http://www.academierainier3.mc/fr/electroacoustique/monaco-electroacoustique-2017

[2] https://github.com/LeugimSan/DETMUS

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

Nuevo entorno para Big Data con R: `sparklyr`

Aurora Baluja González[1,2,3]
Javier López Cacheiro[4]

[1]S. de Anestesioloxía e Reanimación. Hospital Clínico Universitario. Santiago de Compostela.
[2]Dept. de Cirurxía e Especialidades Médico-cirúrxicas. Universidade de Santiago de Compostela.
[3]Instituto de Investigacións Sanitarias (IDIS). Santiago de Compostela.
[4]Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia (CESGA).

RESUMEN

La tecnología que permite la lectura y escritura de macrodatos ("Big Data") ha experimentado un espectacular avance en los últimos años, con la aparición de plataformas open source. El paquete `sparklyr`, lanzado en 2016 por RStudio, permite una comunicación directa con la API de Spark para "dataframes2", utilizando la sintaxis de R.

Palabras y frases clave: Big Data, Spark, R, Scala, mapreduce, SQL.

1. INTRODUCCIÓN

La tecnología que permite la lectura y escritura paralela y eficiente de macrodatos ("Big Data") ha experimentado un espectacular avance en los últimos años, con la aparición de plataformas open source -Hadoop, Spark-, y lenguajes como Scala.
El paquete `sparklyr`, lanzado en 2016 por RStudio, permite una comunicación directa con la API de Spark para "dataframes2", utilizando la sintaxis de R.

2. MATERIAL Y MÉTODOS

El paquete `sparklyr` está desarrollado por RStudio, y establece una interfaz directa para el manejo de tablas estructuradas (dataframes) de Spark-SQL. Utiliza la sintaxis familiar de R para:

- Abrir una conexión con el contexto Spark, que a su vez puede correr en local o sobre un clúster Hadoop.
- Realizar operaciones vectorizadas por debajo de la interfaz, en Spark.
- Permite utilizar algunas funciones familiares del paquete `dplyr`, y funciones de más bajo nivel para operar con los dataframes. Permite incluso queries directamente en lenguaje SQL.
- Una vez procesados los datos, se pueden recoger los resultados en el entorno nativo de R para gráficos, tablas, etc.
- Permite integrarse con algoritmos de las librerías de machine learning de Spark (MLlib) y de H2O.

Además, presenta las siguientes diferencias sobre el paquete `sparkR`, que fue la primera interfaz de R con Spark:

- `sparklyr` no permite emplear UDF (funciones definidas por el usuario) arbitrarias. En su mayor parte son funciones de `dplyr` traducidas a SQL.
- Sin embargo, `sparklyr` presenta una mayor estabilidad e integración en las distintas versiones de Spark.
- `sparklyr` también tiene mejor acceso a la libería MLlib de Spark.

3. CONCLUSIONES

El paquete `sparklyr` es una excelente interfaz para el manejo de ficheros de datos en el entorno Spark.
Basándonos en nuestra experiencia, el paquete `sparkR` es bastante más inestable en cuanto a funcionamiento que `sparklyr`.

Referencias

[1] Spark SQL, DataFrames and Datasets Guide. https://spark.apache.org/docs/latest/sql-programming-guide.html.

[2] `sparklyr`: R interface for Apache Spark. http://spark.rstudio.com/index.html.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

R EN PARALELO E HPC

Diego Mairena[1], Andrés Gómez[1], Aurelio Rodríguez[3] e Santiago Cerviño[4]

[1]Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia (CESGA)
[2]Instituto Español de Oceanografía Centro Oceanográfico de Vigo

RESUMO

O emprego de R en paralelo permite unha redución considerable do tempo de execución da maioría dos programas. Se combinamos isto cos recursos dun centro HPC, podemos obter un gran rendemento, tanto na execución de scripts propios de R, como no emprego de certas aplicacións que fan uso de R.

Palabras e frases chave: HPC, R en paralelo, Gadget.

1. INTRODUCIÓN

O soporte de paquetes para o emprego de R en paralelo comezou ca versión 2.14.0[1], na que se incluía o paquete "parallel", incorporando copias dos paquetes "multicore" e "snow". Dende entón, os paquetes para o uso de R en centros HPC foron medrando ata o día de hoxe, onde existen tanto paquetes de paralelismo explícito ("foreach", "snow") e implícito ("parallel", "Rdsm"), permitindo aproveitar ó máximo os recursos dispoñibles neste tipo de centros.
A versatilidade de R permite que poida adaptarse a calquera entorno computacional, de aí que apareceran paquetes para o uso de R en paralelo relacionados con big data ("RHIPE", que proporciona unha interface entra Hadoop e R), computación Grid, uso de GPUs, ou incluso paquetes para a integración de R cos distintos sistemas de colas empregados en centros HPC (por exemplo, "rslurm").
Ó mesmo tempo, son moitas as aplicacións que fan uso de R para a súa execución, incluso ata o punto de ter paquetes específicos para algunhas delas. Veremos o caso de Gadget[2] (un modelo de ecosistemas mariños), onde o paquete Rgadget proporciona unha integración completa ca propia aplicación, permitindo incluso a súa execución en paralelo.

2. PAQUETES PARA O EMPREGO DE R EN PARALELO

Á hora de empregar R nun entorno HPC, hai que ter en conta tanto os recursos dispoñibles como as limitacións dos propios paquetes. Paquetes como "doParallel" ou "foreach", permiten a execución de R nun entorno multicore, pero sempre dentro do mesmo nodo de cálculo, xa que non están deseñados para facer uso de MPI. O uso destes paquetes permiten unha paralelización moi simple de certas estruturas de programación, como poden ser os bucles tipo "for".
Sen embargo, paquetes como "snow" ou "Rmpi", permiten o uso de varios nodos de cálculo, o que aumenta considerablemente os recursos dispoñibles. Estes paquetes necesitan a maiores dunha ferramenta MPI que permita o envío das mensaxes entre os diferentes nodos, e nalgúns casos, a programación empregando estes paquetes pode ser lixeiramente máis complicada ca no caso de paquetes que empregan o entorno multicore.
Ó mesmo tempo, hai que ter en conta a integración dos diferentes paquetes co sistema de colas que se empregue no centro HPC. Paquetes como "parallel" ou "doParallel", crean o seu propio entorno paralelo unha vez comeza a execución do programa, e o sistema de colas limítase a xestionar os recursos solicitados para o traballo. Un exemplo básico do paquete "doParallel" combinado co paquete "foreach" é a paralelización dunha estrutura tipo "for":

```
library(doParallel)
registerDoParallel(cores=as.numeric(Sys.getenv(\SLURM_NTASKS")))
x <- iris [which(iris[,5] != \setosa"), c(1,5)]
parallel_time <- system.time({
r <- foreach(icount(100000), .combine=cbind) \%dopar\% {
ind <- sample(100, 100, replace=TRUE)
result1 <- glm(x[ind,2]~x[ind,1], family=binomial(logit))
coefficients(result1)}})
parallel_time
```

Os resultados do tempo necesario para a execución deste script, en función dos números de cores empregados, poden verse na Figura 1. Pode observarse unha rápida diminución do tempo ao pasar de 1 a 5 cores, sen embargo, chégase a unha estabilización ao superar os 12 cores, no que o tempo de execución permanece case constante ata os 24 cores.

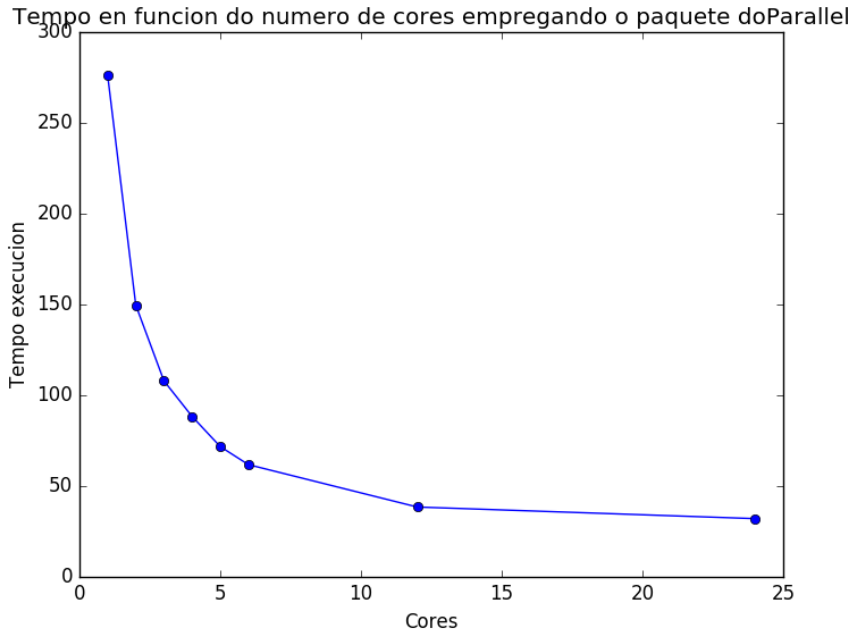


Figura 1: Variación do tempo de execución en función dos cores empregados ao utilizar o paquete doParallel.

Tamén existen certas funcións de R (como por exemplo as relacionadas ca álxebra lineal) que fan uso dun paralelismo implícito dado polas librerías matemáticas empregadas para a compilación de R (MKL, OpenBlas, etc...). No seguinte exemplo, unha multiplicación de matrices escala de forma considerable en función dos recursos dispoñibles para a súa execución:

```
N <- 10000
system.time(x<- matrix(rnorm(N^2), N, N) %*% matrix(rnorm(N^2), N, N))
```

Neste caso, non se fai ningunha referencia explícita ó número de cores a empregar como no caso anterior, é a propio librería a encargada de paralelizar a operación en función dos recursos que ten dispoñibles. Os resultados de tempo en función do número de cores poden verse na Figura 2, onde pode observarse un comportamento similar ó do paquete "doParallel", onde a partir dos 6 cores a curva vaise aproximando a unha asíntota.
Por outra banda, os paquetes como "Rmpi", fan uso de funcións específicas como "mpi.spawn.Rslaves" ou "mpi.bcast.cmd", que á súa vez dependen da ferramenta MPI que se empregue, e o sistema de colas ten que soportar o uso deste tipo de funcións.

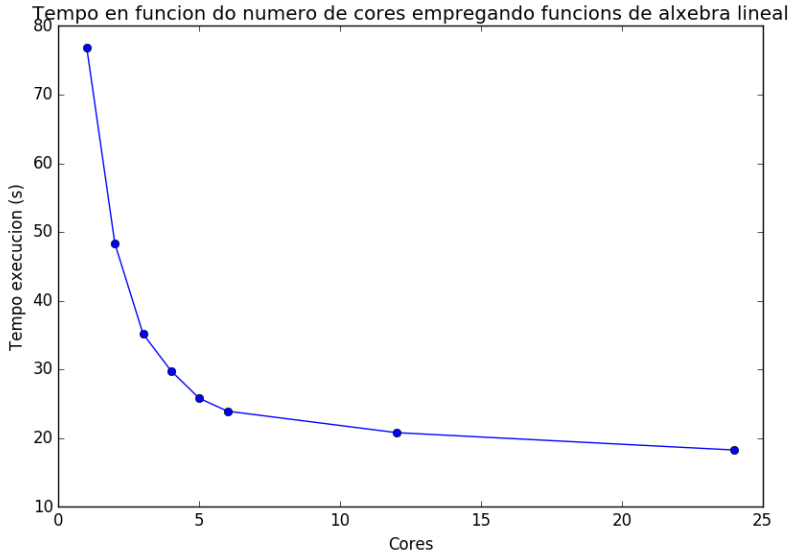


Figura 2: Variación do tempo de execución en función dos cores empregados na multiplicación de matrices.

Un exemplo claro deste caso pode verse con Slurm e os paquetes "Rmpi" e "snow". Slurm emprega o comando "srun" para facer o envío de traballos en paralelo e facer a integración cos recursos solicitados. Este comando fai unha expansión de procesos que non está soportada por estos paquetes de R, polo que non é posible executar o programa. Sen embargo, co comando "mpirun‘" (que proporcionan as ferramentas MPI), si é posible expandir os procesos de forma correcta, aínda cun pequeno matiz: independentemente dos procesos que necesitemos expandir ou os recursos solicitados, sempre é necesario lanzar "mpirun -np 1".
Un exemplo simple do paquete "Rmpi" é detectar os distintos procesos e o host no que que se executan:

```
library(Rmpi)
ns <- mpi.univese.size() -1
mpi.spawn.Rslaves(nslaves=ns)
mpi.bcast.cmd( id <- mpi.comm.rank() )
mpi.bcast.cmd( ns <- mpi.comm.size() )
mpi.bcast.cmd( host <- mpi.get.processor.name() )
mpi.remote.exec(paste("I am", id, "of", ns, "running on", host))
mpi.close.Rslaves(dellog = FALSE)
mpi.exit()
```

Como pode verse, a sintaxe é moi similar ás funcións propias de ferramentas MPI, e a programación empregando este tipo de paquetes aseméllase bastante á empregada por outras linguaxes como C ou Fortran no paradigma de MPI.
Por último, o emprego de certos paquetes de R desenvoltos especificamente para certas aplicacións proporcionan una gran liberdade á hora da súa execución. Gadget[2] é unha aplicación que realiza simulacións estadísticas de ecosistemas mariños, e ó mesmo tempo, emprega diversos paquetes como "mfdb" (para a xestión dos datasets empregados para estimar os parámetros do modelo) ou Rgadget, que proporciona un conxunto de utilidades e análises específicos para a execución de Gadget. Aínda que esta aplicación en principio non está deseñada para a súa execución en paralelo, empregando R podemos paralelizar certas partes e reducir considerablemente o tempo de execución.
Ó mesmo tempo, non só podemos paralelizar as propias funcións de R, se non que tamén podemos executar varias copias de Gadget, en función do número de cores dispoñibles. Neste caso, empre-

gando o paquete "parallel", podemos crear un cluster cos cores solicitados, e en cada un deles, teremos acceso ós ficheiros necesarios para a execución, así como o propio Gadget:

```
ncores<-as.numeric(Sys.getenv("SLURM_NTASKS"))
cluster <- makeCluster(ncores,type="FORK")
clusterExport(cluster,c("input.file"))
clusterExport(cluster,c("gadget.exe"))
parSapply(cluster,ncases,gadget.boot.execute, ...)
```

A execución de cada un destes subprocesos realízase empregando a función parSapply, unha versión paralela da función "sappply", que permite a execución simultánea dunha función un número determinado de veces, en función do tamaño do cluster. En conxunto, todo isto permite executar múltiples procesos simultáneos de Gadget simplemente executando un script de R.

3. CONCLUSIÓNS

Como puidemos comprobar, o emprego de paquetes para a paralelización de R permite unha redución considerable de tempo á hora de executar, e se se empregan recursos dun centro HPC, é necesario coñecer as limitacións dos paquetes cando se combinan co sistema de colas dispoñible. E aínda que esta redución é considerable, tamén se pode observar unha estabilización do tempo cando se chega a certo valor do número de cores empregados.
Por último, o emprego de paquetes específicos para certas aplicacións, permiten a execución das mesmas en paralelo, sen necesidade dunha excesiva programación adicional, e sempre sen saír do entorno de R.

Referencias

[1] CRAN Task View: High-Performance and Parallel Computing with R. Dispoñible en: https://cran.r-project.org/web/views/HighPerformanceComputing.html. Última visitia 10 Agosto 2017

[2] Gadget. Dispoñible en: https://github.com/Hafro/gadget. Última visitia 10 Agosto 2017

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

R COMO PIEDRA ANGULAR DE PROYECTOS DE DATA SCIENCE

Daniel Prieto Rodríguez[1]

[1]Minsait by Indra

RESUMEN

Los proyectos de Data Science se dividen en varias fases que van desde los estadíos iniciales de diseño de la modelización hasta la puesta en producción. En esta presentación mostraremos como R convive con el resto de elementos del ecosistema de este tipo de proyectos en cada una de las fases. En otras palabras, se verá una pequeña aproximación al uso aplicado de R dentro del ecosistema de herramientas, en la realidad empresarial de una industria de distribución.
En primera instancia, librerías tales como "RODBC", "RJDBC", "sparklyr" y "dplyr" permiten, por un lado conectarnos a distintas fuentes de datos y por otro, realizar procesados y análisis exploratorios de los datos. A continuación, "caret" permite evaluar y parametrizar una gran cantidad de modelos de machine learning mientras que, cuando tenemos un volumen de datos muy elevado, "sparklyr"nos proporciona una interfaz de cara a probar los algoritmos de aprendizaje automático incluidos en las librerías de Spark. De forma transversal a todas las fases, la librería "rmarkdown" nos permite generar informes de forma automatizada para contrastar las hipótesis y resultados. Además, este tipo de proyectos se engloban en librerías propias de R en las que se hace uso de programación orientada a objetos y patrones de diseño; a fin de llevar a cabo la implementación de algoritmos de propiedad industrial.
Finalmente, compartiremos la necesidad de apoyar los desarrollos en R sobre otras herramientas y plataformas que se requieren para conseguir la puesta en producción de proyectos empresariales.

Palabras y frases clave: Data Science, R, Spark, sparklyr, distribución.

Referencias

[1] JJ Allaire, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman, and Ruben Arslan. *rmarkdown: Dynamic Documents for R*, 2017. R package version 1.5.

[2] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-76.

[3] Javier Luraschi, Kevin Ushey, JJ Allaire, and The Apache Software Foundation. *sparklyr: R Interface to Apache Spark*, 2017. R package version 0.6.1.

[4] Brian Ripley and Michael Lapsley. *RODBC: ODBC Database Access*, 2017. R package version 1.3-15.

[5] Simon Urbanek. *RJDBC: Provides access to databases through the JDBC interface*, 2014. R package version 0.2-5.

[6] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2016. R package version 0.5.0.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

PREDICCIÓN DE SERIES TEMPORALES EN DATASETS MULTIDIMENSIONALES DE SITUACIONES DE NEGOCIO MEDIANTE PARALELIZACIÓN AGRADABLE EN R SERVER

Antonio Soto[1]

[1]SolidQ

RESUMEN

La aplicación de series temporales a la predicción de valores en situaciones de negocio está ampliamente extendida. Técnicas como regresiones multivariantes y autoregresiones vectoriales son usadas ampliamente, pero métodos más novedosos propios de la minería de datos, como SVMs, boosted trees y redes neuronales, se basan en modelos de tipo consultivos los cuales, a la hora de ser usados para predicciones de tipo 1-paso adelante, carecen de datos para poder ser alimentados y pronosticar la o las variables dependientes de interés. Este tipo de problemas se presentan en predicciones temporales en situaciones de negocio que suelen tener un valor específico, cómo puede ser el número de reservas de hotel / piso en plataformas en línea, asociadas a un conjunto de datos multidimensional, como geografía de origen y geografía destino, perfil de cliente, perfil de hotel, duración de estancia, etc., siempre asociadas a un eje temporal.
Ante este problema, hemos aplicado distintos modelos predictivos de series temporales para predecir a un horizonte finito un determinado valor haciendo n-particiones inconexas altamente paralelizables del conjunto de datos disponible con el uso de R Server, y generando una serie temporal única para cada partición. Específicamente se predice el número de recogidas de usuarios de taxi en la ciudad de Nueva York, por origen, destino, tipo de pago, duración del trayecto y otras dimensiones, pero el problema es fácilmente trasladado a otras situaciones de negocio. El conjunto de todas las series temporales es luego reagrupado generando una tabla multidimensional de futuros con valores predichos, distribuciones de las variables independientes coherentes con los valores históricos y ejecutado todo dentro del entorno paralelizado de R Server.
Nuestra aproximación resuelve un problema específico extenso dentro de la minería de datos, al mismo tiempo que mediante la utilización de R Server reducimos tiempos entrenamiento y predicción respecto a soluciones similares desarrolladas con foreach y paquetes de paralelización de R (snow, future).

Palabras e frases clave: Series Temporales, Minería de Datos, Machine Learning, Microsoft R Server.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

ANÁLISE DA INCIDENCIA DA LEUCEMIA GRANULOCÍTICA EMPREGANDO A ESTIMACIÓN NON PARAMÉTRICA DE CONXUNTOS DE NIVEL

Paula Saavedra-Nieves[1] e Alberto Rodríguez-Casal[2]

[1]Universidade de Vigo
[2]Universidade de Santiago de Compostela

RESUMO

Un problema fundamental en epidemioloxía é determinar se as zonas nas que se concentran os casos dunha enfermidade se corresponden coas rexións máis poboadas ou, polo contrario, existen áreas de risco nas que a súa incidencia é maior. A estimación non paramétrica de conxuntos de nivel é unha ferramenta estatística útil para abordar este tipo de cuestións. A partir dun conxunto de datos reais, analizaremos a incidencia da leucemia granulocítica en Lancashire e Greater Manchester estimando conxuntos de nivel con R.

Palabras e frases chave: Estimación non paramétrica. Conxuntos de nivel. Parámetro de suavizado. Leucemia granulocítica

RESUMO AMPLIADO

A incidencia de certas enfermidades varía dunhas rexións a outras debido á influencia de factores ambientais de risco. Na actualidade, este tipo de cuestións poden ser estudadas a partir das cantidades masivas de datos epidemiolóxicos dispoñibles. A Figura 1 contén 1221 pares de puntos sobre as rexións de Lancashire e Greater Manchester. Correspóndense coas coordenadas xeográficas de residencia para 233 casos de leucemia granulocítica diagnosticados entre 1982 e 1998, xunto con 988 controis. Para unha descrición detallada dos datos, ver Henderson et al. (2002).

Figura 1: Rexións de Lancashire e Greater Manchester sobre o Noroeste de Inglaterra (esquerda), distribución de 233 casos de leucemia diagnosticada (centro) e 988 controis (dereita) sobre as dúas rexións de interese.

Un problema fundamental en epidemioloxía é determinar se as zonas nas que se concentran, por exemplo, os casos de leucemia correspóndense coas zonas máis poboadas ou se existen zonas nas que

a incidencia da enfermidade é maior. A estimación non paramétrica de conxuntos de nivel é unha ferramenta estatística moi útil neste contexto. Esta teoría ocúpase de reconstruír os conxuntos $G(t) = \{x : f(x) \geq t\}$ a partir dunha mostra aleatoria $X_1, \ldots, X_n$ dun vector aleatorio X, onde f representa a función de densidade de X e t, un nivel positivo.
Os estimadores do conxunto de nivel G(t) son fexibles debido a que non se asume ningunha hipótese paramétrica sobre a distribución dos datos. A cambio, certos parámetros de suavizado que non resultan tan determinantes nas reconstrucións dos conxuntos deben ser estimados. Ver Samworth e Wand (2010) ou Walther (1997).

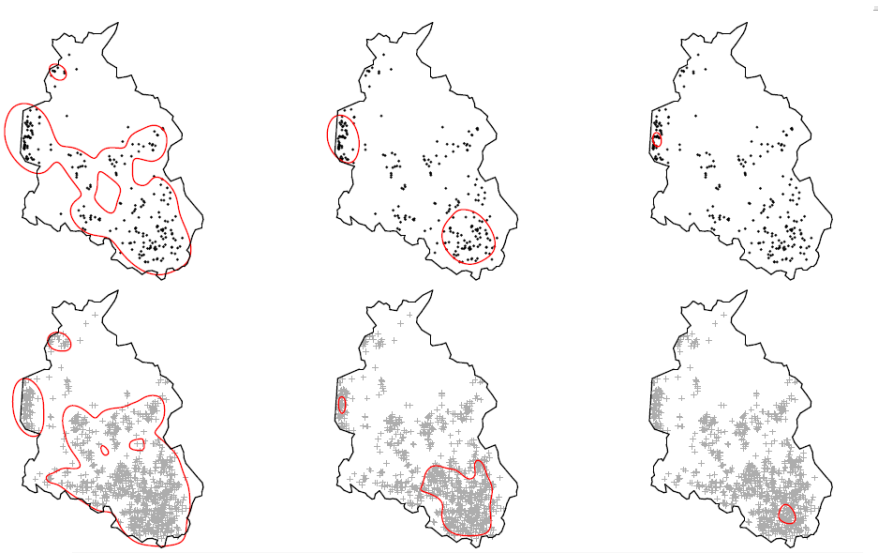


Figura 2: Estimadores non paramétricos dos conxuntos de nivel para as mostras de 322 casos de leucemia (primeira fila) e 988 controis (segunda fila). En la cada columna, $t = t_i, i = 1, 2, 3$ (columna i) con $t_1 < t_2 < t_3$.

Neste traballo, empregaremos R para reconstruír conxuntos de nivel a partir das mostras de casos e controis de leucemia granulocítica presentadas previamente. Na Figura 2, móstranse estimacións obtidas considerando distintos valores do nivel t. Dacordo cos resultados mostrados, obsérvase que casos e controis distribúense de xeito claramente distinto, ver terceira columna na Figura 2. Na parte norte do mapa, os casos de leucemia presentan unha intesidade maior. Greater Manchester é unha das rexións metropolitanas máis grandes de Inglaterra. Sen embargo, Lancashire é unha zona industrializada con niveis altos de contaminación que poden ser determinantes na elevada incidencia da enfermidade.

AGRADECEMENTOS

Os autores agradecen o soporte económico do Ministerio de Economía e Competitividade a través do proxecto MTM2016-76969-P (AEI/FEDER, UE) e o soporte económico da Xunta de Galicia a través dos ERDF (Grupos de Referencia Competitiva) ED431C 2016-040.

Referencias

[1] Henderson R., Shimakura S., Gorst D. (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association* 97, 965-972.

[2] Samworth R.J., Wand M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Annals of Statistics* 38, 1767-1792.

[3] Walther G. (1997). Granulometric smoothing. *Annals of Statistics* 25, 2273-2299.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

RStudio COMO HERRAMIENTA PARA DESARROLLO DE MATERIAL DOCENTE INTERACTIVO Y MULTIMEDIA

Alejandro Quintela del Río[1]

[1]Departamento de Matemáticas, Universidad de A Coruña

RESUMEN

Además de servir para la realización de programas y material científico reproducible, se mostrará la versatilidad de Rstudio para la creación de material docente (clases y apuntes). En ellos pueden incluirse elementos multimedia (videos, animaciones, tablas y gráficos interactivos), código html, código latex, y facilitar el proceso de aprendizaje creando tutoriales y preguntas. Al poder convertirse directamente en una página web, los apuntes serán accesibles a través de internet a los alumnos, además de poder publicarse en pdf o word.

Palabras y frases clave: R, Rstudio, Rmarkdown, html.

1. INTRODUCIÓN

R y **Rstudio**, junto con el paquete **Rmarkdown**[1], permiten producir documentos que pueden convertirse fácilmente en un documento Word, PDF, o guardarse como un archivo HTML que se puede alojar en cualquier sitio web. El archivo .Rmd (Rmarkdown) puede contener sólo texto, como un simple informe escrito, o un documento mucho más complejo, con código R para producir diagramas, mapas o cualquier gráfico que se pueda generar mediante R. No es necesario producir los gráficos y las tablas, y luego añadirlos a un documento word, odt o latex, sino que todo ese trabajo puede ser realizado de una sola vez. Si se requiere que el documento se reproduzca con nuevos datos, resulta muy sencillo actualizar las instrucciones para volver a ejecutar el código y producir una nueva versión del documento.
Actualmente, estamos trabajando en la realización de unos apuntes genéricos de una asignatura de estadística básica (descriptiva, probabilidad, variables aleatorias e introducción a la inferencia) utilizando estas herramientas, y paulatinamente vamos descubriendo lo que parece un sinfín de posibilidades.

2. CONTENIDOS BÁSICOS

Una de las grandes ventajas de Rmarkdown es la producción de documentos formateados mediante una sintáxis muy simple, la cual puede consultarse en la plantilla general de referencia (Cheat Sheet)[2].
Mediante el paquete **knitr**[3] es posible introducir trozos de código (*chunks*) que se ejecutarán al producirse el documento (tanto en word, pdf o html), pudiendo cambiarlo cuando queramos y producir resultados iguales o diferentes.
Al ejecutar estos trozos de código (que incluso pueden incluirse en medio de una línea del documento, generando en él los resultados), produciremos gráficas, tablas de resultados o diagramas que automáticamente se incluirán en el trabajo final.

3. ALGUNAS CARACTERÍSTICAS DE INTERÉS

Inclusión de gráficos:
Es posible introducir gráficos o imágenes generadas en otros programas, con tal de que estén en formato jpg, png o gif, simplemente mediante el código:
`![] (carpeta/nombredefichero)`
Inclusión de videos o gráficos interactivos:
Mediante el formato

```
<iframe width="640" height="360"

src="https://www.youtube.com/embed/-Zu93vbId-0I"

frameborder="0" allowfullscreen></iframe>
```

insertaremos un video, en este caso directamente desde youtube. Para conseguir la dirección (en rojo), tendremos que ir a youtube (o la plataforma correspondiente y buscar la opción "insertar" para obtener la url correspondiente).

Inclusión de tablas interactivas:
Con el siguiente trozo de código (chunk) leemos directamente un fichero excel y lo convertimos en una tabla interactiva (obviamente solo en formato html)

```
library(readxl)

Titanic <- read_excel("Pasajeros-Titanic.xlsx")

datatable(Titanic, options = list(pageLength = 5)) # Interactive table
```

Inclusión de código latex:
Escribiendo $\$a + bx\$$ escribimos esta ecuación en línea, y $\$\$a + bx\$\$$ una ecuación de una línea entera.

Inclusión de código html:
La sintaxis de Rmarkdown está bastante limitada en cuanto a posiblidades de formateo (* para cursiva ** para negrita y `` para recuadrar). Sin embargo, se puede incluir código html al principio del documento para obtener resultados de formateo "al gusto". Por ejemplo , con estas líneas al principio del documento Rmd definimos tres nuevos estilos: estilo **ejemplo** (tamaño 16, en negrita y tipo de letra Lucida Console en color azul), y estilo **resaltar** y estilo **borde**:

```
<style type="text/css">

.chart-important {  /* chart_title  */
   font-size: 10px;
   font-family: Algerian;
   font-color:red;
}
ejemplo
{
font-size: 16px;
font-weight:bold;
font-family:Lucida Console;
color: blue;
}
 resaltar{
background-color:#ffff00;
}
```

```
 borde{
      border: 1px solid red;
    }
</style>
```

Para aplicarlo, escribimos:
`<ejemplo>cualquier texto </ejemplo>`.
Equivalentemente, `<resaltar>cualquier texto </resaltar>`, o `<borde>cualquier texto </borde>` y `cualquier texto` aparecerá con el formato que hayamos creado.
Las posibilidades para crear formatos con html son tantas como queramos, y para obtener el código html para un formato que nos interese, basta con ir a una de las muchas páginas que hay en internet de edición html, por ejemplo [4].

Realización de tutoriales y preguntas interactivas
El Rstudio permite directamente, a través del paquete **manipulate**, la realización de gráficos interactivos, o bien a través de los paquetes **plotly** y **ggplot2** mediante la orden `ggplotly`. Además de esto, mediante la inclusión del paquete **learnr**[5], es posible crear de forma sencilla tutoriales y ejercicios.
Pueden verse (casi) todas las opciones aquí comentadas y muchas más en el libro de Yihui, X. [6].

Referencias

[1] http://rpubs.com/marschmi/RMarkdown/

[2] https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-spanish.pdf

[3] Xie, Yihui. 2015. Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.

[4] http://html-color-codes.info/html-editor/

[5] https://rstudio.github.io/learnr/

[6] Xie, Yihui. 2017. Bookdown: Authoring Books and Technical Documents with R Markdown. https://bookdown.org/yihui/bookdown/

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

DIBUJANDO CURVAS ROC EN R

Arís Fanjul Hevia[1]

[1]Dpt. de Estadística, Análisis Matemático y Optimización, Universidade de Santiago de Compostela

RESUMEN

La curva ROC (del inglés Receiver Operating Characteristic curve) es una herramienta estadística que se emplea para evaluar un sistema de clasificación [1]. Aparece sobre todo en el entorno médico, donde se utiliza para determinar cómo de bueno es un test de diagnosis en el que una variable es utilizada para clasificar a la población en sanos y enfermos. Además, tiene aplicaciones en otras áreas como la psicología, las finanzas, la meteorología y, en general, en cualquier campo en el que interese poder medir la capacidad discriminativa de un sistema de clasificación que sirva de base para una toma de decisiones.

Existe una gran variedad de librerías en R que permiten representar este tipo de curvas. En algunas de esas librerías solo aparecen como complemento para otro tipo de estudios, mientras que existen otras como pROC [2], ROCR [3] o OptimalCutpoints [4] dedicadas exclusivamente a la construcción y análisis de las curvas ROC. También destacan algunas centradas en el análisis de supervivencia como survivalROC[5], ya que, como se ha mencionado, es una herramienta especialmente empleada en el ámbito biomédico.

El objetivo de esta charla es explicar brevemente qué librerías existen para tratar las curvas ROC y cuáles son sus principales funciones. Se hablará de los distintos métodos que hay para estimar estas curvas [6], cómo encontrar los puntos de corte óptimos para la toma de decisiones, cómo calcular regiones de confianza y cómo comparar dos o más de estas curvas. Además, se expondrán problemas con datos reales como ejemplo de aplicación de estas técnicas.

Palabras e frases clave: AUC, curvas ROC, punto de corte óptimo.

AGRADECIMIENTOS

Este trabajo está financiado por el Ministerio de Educación, Cultura y Deporte de España (Beca FPU 2014/05316). También está dentro del proyecto MTM2013-41383-P y MTM2016-76969-P que incluye ayudas del Fondo Europeo de Desenvolvemento Rexional (FEDER) y de la red IAP de StUDyS de la Belgian Science Policy.

Referencias

[1] M.S. Pepe (2003). The statistical evaluation of medical tests for classification and prediction. *Oxford University Press*, New York.

[2] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, M. Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77. DOI: 10.1186/1471-2105-12-77. URL http://www.biomedcentral.com/1471-2105/12/77/.

[3] T. Sing, O. Sander,N. Beerenwinkel, T. Lengauer (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20), 7881. URL: http://rocr.bioinf.mpi-sb.mpg.de.

[4] M. Lopez-Raton, M.X. Rodriguez-Alvarez, C. Cadarso-Suarez, F. Gude-Sampedro (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*, 61(8), 1-36. URL http://www.jstatsoft.org/v61/i08/.

[5] P.J. Heagerty, P. Saha-Chaudhuri (2013). survivalROC: Time-dependent ROC curve estimation from censored survival data. R package version 1.0.3. https://CRAN.R-project.org/package=survivalROC.

[6] L. Gonçalves, A. Subtil, M.R. Oliveira, P. de Zea Bermúdez. (2014). *Statistical Journal* 12, 1–20.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

A ECOLOXÍA NA SÚA REVOLUCIÓN DE CANTIDADE MASIVA DE DATOS

Miguel Branco[1]

[1]Universidade de Vigo

RESUMO

Os estudos de biodiversidade a escala global e os cambios nas guías editoriais e planos gobernamentais, que obrigan á publicación de datos, xeraron na última década grandes fontes de información ligada á ecoloxía. Para o seu aproveitamento, creáronse librarías en R que cobren tanto os pasos de obtención, integración e meta-análise dos datos como a súa publicación. Así, librarías como "taxize" e "phyloseq" permiten a obtención de censos de organismos, outras como "rfigshare" a publicación de estudos de campo e outras como "aRxiv" a publicación de artigos, froito dos estudos.

Palabras e frases chave: ecoloxía, reproducibilidade, datos abertos, rfigshare, taxize, phyloseq.

1. INTRODUCIÓN

A ecoloxía está a vivir un momento de masiva publicación de datos de campo, de proxectos tanto locais como globais, e que esixen un cambio na dinámica de traballo nesta disciplina. Arestora estanse a depositar os datos ligados a estes estudos en repositorios públicos e con licenzas libres, permitindo o seu uso entre múltiples grupos de traballo. Existen repositorios tanto xeneralistas e que acollen a datos de estudos ecolóxicos de múltiples tipos, como *Figshare* ou *Dryad*; como específicos para información taxonómica e de abundancias de especies, como o *NCBI Taxonomy ou GBIF*. R conta arestora cun ecosistema de paquetes que permite abordar todo o proceso de obtención, manipulación e almacenamento desta información de biodiversidade; así como a súa publicación e a súa análise ecolóxica. R está a facilitar a reproducibilidade dos estudos e a colaboración en grupos globais na ecoloxía, poñéndoa na súa nova década.

2. A ECOLOXÍA E OS PROXECTOS GLOBAIS

A investigación na ecoloxía, e campos ligados como a evolutiva, está a pasar por un momento dunha crecente produción de datos, que cada vez son máis variados na súa forma e na súa orixe. Compílanse continuamente, como exemplo, datos físicos e bioxeográficos de especies de fitoplancto mariño, metaxenomas de poboacións de bacterias da pel de mamíferos ou a composición do solo para a estudos de edafoloxía. Os estudos na ecoloxía comezaron na década de 1960 a ser da "gran ecoloxía" *[big ecology]*: moitos dos estudos locais, de pouco custo, fóronse convertendo en estudos de escala global ou de longa escala temporal, coordinados por proxectos internacionais e multi-gobernamentais. Proxectos internacionais como a *Global Biodiversity Information Facility (GBIF)*, o *National Ecological Observatory Network (NEON)* ou a *Ocean Observatories Initiative (OOI)* marcáronse retos de catalogación da biodiversidade global ou variables bioxeoquímicas. Con isto, convertéronse en grandes fontes de datos para a ecoloxía actual, en particular desde o 2000. Os obxectivos dos estudos ecolóxicos supoñen empregar diferentes tipos de datos ou realizar meta-análises. Empréganse, como exemplo, distintos grupos taxonómicos de especies, datos de secuenciación masiva, ou múltiples puntos de mostraxe en todo o planeta e que proceden de múltiples fontes de datos.

Na última década a crecente implicación dos ecólogos en grupos interdisciplinares de proxectos globais fomentou a dinámica de compartir as observacións nesas bases de datos, e co cal, a súa reutilización. Arestora, tanto os comités editoriais das revistas, como os comités gobernamentais incentivan aos autores a depositar e a compartir os datos. Con iso, pretenden facilitar a reproducibilidade das investigacións, a colaboración e a garantir a verificabilidade dos estudos. Mesmo moitos dos comités de edición de xornais científicos xa establecen como unha obriga a publicacións de datos. Os proxectos gobernamentais como o «Open Science (Open Access)» *[Ciencia Aberta, Acceso Libre]*, da Unión Europea no Horizon 2020, céntranse na liberación de datos e resultados das investigacións financiadas con fondos públicos. Están a recoñecer que os datos creados nas ciencias, como a ecoloxía, se débense compartir, estar accesibles, e así, pertencer á sociedade.

3. DATOS DE BIODIVERSIDADE E R: ALMACENAMENTO, OBTENCIÓN E PROCESADO

Os proxectos globais de estudo foron creando múltiples repositorios de almacenamento. Un dos principais obxectivos de estudo da ecoloxía é catalogar e comprender a biodiversidade global. Para iso son datos de interese a información taxonómica, metaxenomas ou información ambiental dos lugares mostreados. Os principais lugares para a publicación e busca desa información de campo son algúns temáticos como a *Global Biodiversity Information Facility (GBIF), Encyclopedia of Life (EOL)*, a *International Barcode of Life (iBOL)*. Algún outro fai de metabuscador para múltiples fontes, como é *Data Observation Network for Earth (DataONE)*. E outros repositorios son xenéricos e permiten a publicación e obtención de datos diversos tanto en formato como en tipoloxía; como son "Figshare" e "Dryad".

R conta con librarías que automatizan a obtención e procesamento da información almacenada nestes repositorios. "rentrez" outorga acceso a múltiples bases de datos do NCBI, como a GenBank, e "rbgif" a posición taxonómica e abundancias de especies almacenadas en GBIF. Outras librarías máis universais como "taxize" buscan en máis de 13 repositorios a posición taxonómica de listados de especies e almacénanas, lígandoas coas súas posicións taxonómicas e cos identificadores das bases de datos. "phyloseq" fai algo similar só que tratando datos de OTUs de metaxenomas e censándoas. "traits" busca as características fenotípicas das especies de interese en bases de datos do NCBI, Traitbank, ou Birdlife. Outros como rnoaa facilitan a extración de datos ambientais, como a temperatura da auga oceánica que almacena a N*ational Oceanic and Atmospheric Administration (NOAA)*; e "robis" superpón datos de abundancias de especies mariñas obtidos de *Ocean Information Biogeographic System (OBIS)* a información espacial. Ademais de para obtelos, R conta con múltiples librarías que facilitan a súa compartición e publicación. "rfigshare", "rdryad" ou "rdataone" permiten compartir os datos nas bases de datos e outros como "aRxiv" permiten a pre-publicación dos artigos, froito de meta-análises, en arxiv.org.

4. CONCLUSIÓNS

R está a favorecer a publicación e o tratamento de datos na ecoloxía e está converténdose nun dos seus piares. Facilita a obtención e tratamento de datos de diferentes fontes e a preparación para a súa publicación, con finalidades como o estudo da biodiversidade e bioxeografía globais. O ecosistema de librarías de R cobre continuamente novos espazos que se necesitan nos novos retos de estudos globais da ecoloxía.

Referencias

[1] Chamberlain, S. A., & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. F1000Research, 2, 191. http://doi.org/10.12688/f1000research.2-191.v2

[2] Hampton S.E., Anderson S.S., Bagby S.C., Gries C., Han X., Hart E.M., Jones M.B., Lenhardt W.C., MacDonald A., Michener W.K., Mudge J., Pourmokhtarian A., Schildhauer M.P., Woo K.H. and Zimmerman N. (2015) The Tao of open science for ecology. Ecosphere 6(7): 1–13.

[3] Open Science (Open Access). The EU Framework Programme for Research and Innovation. European Union. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access

[4] McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE, 8(4), e61217. http://doi.org/10.1371/journal.pone.0061217

[5] Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biology, 13(11), e1002295. http://doi.org/10.1371/journal.pbio.1002295

[6] William K. Michener, Ecological data sharing, Ecological Informatics, Volume 29, 2015, Pages 33-44, ISSN 1574-9541, http://dx.doi.org/10.1016/j.ecoinf.2015.06.010.

IV Xornada de Usuarios de R en Galicia
Santiago de Compostela, 19 de outubro do 2017

EXTRACCIÓN DE DATOS DA WEB (Web Scraping) CON R

José Luis Juncal Pérez

RESUMO

Extracción de datos a un ficheiro CSV, con formato definido polo cliente, de diversas webs abertas ó público.

Palabras e frases chave: obtención de datos, scraping web, rvest, tidyverse.

1. INTRODUCIÓN

Exporánse as capacidades de extracción de datos da web en R.
Comezaría co porqué da elección dos paquetes utilizados (rvest) e a súa adaptación e integración coa filosofía Tidyverse (http://style.tidyverse.org/).
No campo da extracción de datos na web, fariase un breve repaso ó DOM (http://www.w3.org/DOM/) e farramentas auxiliares para facilitar a localización dos ítems a obter.

2. EXPOSICIÓN

Cada vez máis, necesitamos extraer datos da web. Como analistas, voltámonos máis autónomos no noso traballo xa que non precisamos de táboas ou ficheiros preparados por outros, ou APIs dos servicios web que utilizamos. Logo de evaluar os paquetes dispoñibles (httr, xml2, rvest e Rselenium) decídese utilizar rvest, por tres razóns:

- *potencia*, permite realizar extraccións complexas utilizando CSS Selectors ou XPath,
- *sinxeleza*, non necesita recursos complexos coma no caso de Rselenium,
- *curva de aprendizaxe*, xeito de traballo coherente coa filosofía Tidyverse.

O traballo proposto é a extracción de características das lanas á venda no eCommerce de Katia (https://katia.com/) e adaptalo ó formato de importación de produtos de Prestashop.
Xa que todos os campos a rechear serían repetitivos para a sesión, só se comentarían os máis singulares, con trucos e boas prácticas. Principalmente, trataríamos cómo obter cada característica por separado ou varias á vez, e ver cal modo convén en cada situación. Cómo gardala información, e que campos a maiores se poderían gardar en previsión de usos futuros. Por último, boas prácticas á hora de colleitar información en webs alleas.
Tamén comentaríase por riba as capacidades avanzadas que teriamos con Rselenium, xa que sería o referente á hora de facer traballos máis complexos.

3. CONCLUSIÓNS

Cada vez máis, necesitamos obter datos por nos mesmos. Depender de outros para a obtención e preparación dos datos, pode supor semás (experiencia propia) ou que nunca nolos fagan dispoñibles de xeito adecuado (ficheiro, APIs, etc). Para un analista, dispor libremente de grande cantidade de datos é un dos maiores apoios ó seu traballo, por iso creo que o campo da obtención de datos é máis importante que o post-proceso e limpeza de datos en ficheiros alleos.

AUTORES

IV XORNADA DE USUARIOS DE R EN GALICIA

> ORGANIZA

CIXUG

> COLABORA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA