V XORNADA DE USUARIOS DE R EN GALICIA

| 25 de outubro de 2018
| Santiago de Compostela

LIBRO DE RESUMOS

> ORGANIZA

CIXUG

> COLABORAN

R-Galicia

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE MATEMÁTICAS

> PATROCINAN

amtega Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

V XORNADA DE USUARIOS R EN GALICIA

PROGRAMA
E RESUMOS

Santiago de Compostela

25 de Outubro de 2018

ORGANIZA:

Oficina de Software Libre (OSL) do CIXUG

Editora: Mª José Ginzo Villamayor

ISBN: 9788409058051

PRESENTACIÓN

A Oficina de Software Libre (OSL) do CIXUG comprácese en presentar a V 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 dezaseta relatorios ao longo de todo o día, ademais de dous obradoiros: Introdución a R e Iniciación ó Big Data con R coa libraría sparklyr.

Entre os participantes figuran especialistas do Instituto Galego de Estatística, da Consellaría de Sanidade e outros organismos da Xunta de Galicia, das tres universidades galegas, do Cesga, de ITMATI de Gradiant e de empresas como FINSA, ABanca ou Improving Metrics.

Todo isto non sería posible sen o patrocinio de AMTEGA e a colaboración da Asociación de Usuarios de Software Libre da Terra de Melide (MeLiSA) e 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 2018

O Comité Organizador

COMITÉ ORGANIZADOR

Mª José Ginzo Villamayor
Universidade de Santiago de Compostela

Rafael Rodríguez Gayoso
Concello de Santiago de Compostela

Xabier Sánchez Santos
Consorcio Interuniversitario CIXUG

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

25 de outubro de 2018

ACCESO WIFI NA SEDE

SSID: xornadar

Login: xornadar

Password: 45Hmp%sR

CERTIFICADOS

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

UBICACIÓNS NA FACULTADE

Relatorios: Aula Magna (nivel 3)

Obradoiros: Aulas 1 e 5 (nivel 1).

Cafés: corredor nivel 3.

PROGRAMA

9:25 -	9:50	¿Qué coche de ocasión me compro? Una aproximación con R *Antonio Vidal Vidal (FINSA)*
9:50 -	10:15	R en la generación de informes automáticos a través de la API de Google Analytics *Miguel Boubeta Martínez (Improving Metrics)*
10:15 -	10:40	Planificación estratégica con R *Belén M. Fernández de Castro e Teresa Veiga Rodríguez (ABANCA)*
10:40 -	11:05	Visualización da información estatística utilizando R-Shiny e R-Markdown *Noa Veiguela Fernández (IGE)*
11:35 -	12:00	Visualización interactiva de datos de saúde (creación de dashboards con Shiny) *Miguel Ángel Rodríguez Muiños (Consellería de Sanidade)*
12:00 -	12:25	R en la administración pública. Cálculos con intervalos de fechas *Marcos Fernández Arias* (Amtega - *Xunta de Galicia*)
12:25 -	12:50	The importance of testing R code *Nora Martínez Villanueva (Gradiant)*
12:50 -	13:10	Integración de R en QGIS *Ana Belén Buide Carballosa (Instituto Tecnolóxico de Matemática Industrial - ITMATI)*
13:10 -	13:30	Algoritmo de eficiencia nas descargas de auga *Manuel Antonio Novo Pérez (Instituto Tecnolóxico de Matemática Industrial - ITMATI)*
13:30 -	13:55	R y HPC (uso de R en el CESGA) *Aurelio Rodríguez (Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia - CESGA)*
15:25 -	15:50	Novas librerías para o control estatístico da calidade (QCR) e estudos interlaboratorio (ILS) no contorno da industria 4.0 *Salvador Naya Fernández (UDC)*
15:50 -	16:15	Bookdown: un paquete de R para á creación de libros *Rubén Fernández Casal (UDC) e Tomás Cotos Yañez (UVigo)*
16:15 -	16:40	Estimación de conxuntos con R mediante os paquetes alphahull e alphashape3d *Beatriz Pateiro López (USC)*
16:40 -	17:00	Modelos estadísticos de clasificación con alta dimensión en el número de covariables *Laura Freijeiro González (USC)*
17:00 -	17:20	Comparando métodos diagnósticos en R *Arís Fanjul Hevia (USC)*
17:20 -	17:40	Ferramentas para reducir o tempo de execución en R *Alejandra López Pérez (USC)*
17:40 -	18:00	Unha aplicación Shiny de R para a xestión de recursos en incendios forestais *Mª José Ginzo Villamayor (USC)*
18:00 -	20:00	Obradoiro: Introdución a R *Mª José Ginzo Villamayor (USC)*
		Obradoiro: Iniciación ó Big Data con R coa libraría sparklyr *Aurora Baluja González (CHUS)*

Índice

¿QUÉ COCHE DE OCASIÓN ME COMPRO? UNA APROXIMACIÓN CON R

Antonio Vidal Vidal[1]

[1] Finsa

RESUMEN

En este trabajo se muestra cómo desarrollar un sistema de predicción de los precios de los vehículos de ocasión con los datos que están disponibles en los portales de publicación de anuncios usando únicamente R para realizar todo el proceso.

Palabras y frases clave: Iweb scraping, EDA, machine learning, caso de uso, predicción de precios de venta, ajuste de parámetros.

1. INTRODUCCIÓN

Supongamos que queremos comprar un coche de ocasión. En las páginas de vehículos de ocasión podemos seleccionar el tipo de coche, la marca y el modelo, el combustible y el motor, la antigüedad, el rango de kilómetros e incluso el color y, a continuación, nos muestran los vehículos que cumplen esas condiciones con su precio, pero, ¿cuál es su precio justo?

2. DESARROLLO DEL CASO DE USO

Se puede observar que, aunque las páginas de ofertas de coches de ocasión disponen de todos los datos necesarios, no proporcionan los resultados de la predicción del precio de los coches que aparecen publicados en sus páginas. Sin embargo, una vez que dispones de esos datos, realizar una valoración de los precios de los vehículos es un trabajo muy fácil.

Para demostrarlo, en este caso de uso se han realizado las siguientes acciones con R para generar un sistema que predice el valor de los precios de los vehículos de ocasión con un error MAPE inferior al 10%:

- Capturar los datos de los vehículos seleccionados desde diferentes portales web de listados de ofertas.
- Programar la captura diaria de esos datos.
- Explorar los datos capturados
- Analizar la influencia de diferentes variables en el precio de los vehículos de ocasión
- Generar las variables necesarias para realizar modelos de predicción de los precios, incluyendo el procesado de los textos capturados
- Evaluar diferentes modelos de predicción
- Realizar las predicciones de los precios de los vehículos en venta y analizar la influencia de las variables en esa predicción.

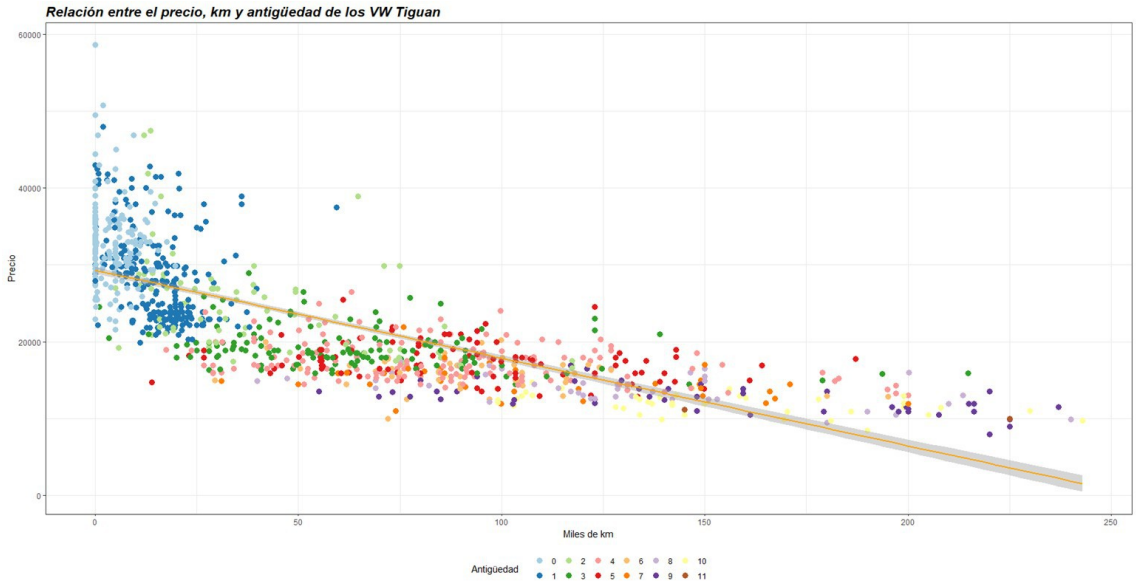


Figura 1: Ejemplo de la relación entre el precio y los miles de km del coche, coloreado en función de la antigüedad.

A lo largo del caso de uso se recomendarán diferentes paquetes de R para realizar cada una de las etapas del proceso, así como los principales retos afrontados en el desarrollo del proyecto.

3. CONCLUSIONES

En este caso de uso se muestra que, a pesar de la importancia del precio en la decisión de compra de un vehículo de ocasión, la mayoría de los portales de anuncios no publican un precio recomendado para un determinado vehículo a pesar de que es relativamente sencillo con R realizar un proceso completo para generar modelos que presentan un error MAPE inferior al 10% sobre el precio publicado para ayudar a los compradores a tomar su decisión.

R EN LA GENERACIÓN DE INFORMES AUTOMÁTICOS A TRAVÉS DE LA API DE Google Analytics

Miguel Boubeta Martínez[1], Eva María González Vior[1], María Elena Naranjo Sánchez[1], José Manuel Pérez Novo[1]

[1] Improving Metrics

RESUMEN

Uno de los objetivos de los analistas digitales es saber cómo presentar de manera fácil los datos que resulten útiles en la toma de decisiones. Esto se transforma en informes con ciertas especificidades que se deben generar periódicamente. Con motivo de la gran cantidad de datos que se recaba en Google Analytics y las limitaciones para manejarlos al gusto del analista, se ha propuesto el uso del software R para la consulta, manejo y generación de informes. Para esto, con las librerías "googleAnalyticsR", "xlsx", "shiny" entre otras, se ha creado una estructura para generar informes con los datos extraídos mediante la API de Google Analytics. En esta estructura de informes se contempla la configuración de diferentes vistas, el manejo de fechas comerciales, configuración a través de R de la tipografía y estilos a imprimir en un informe de tipo Microsoft Excel y finalmente leer y enviar correos con informes adjuntos haciendo uso de la integración de Python y RStudio server. Se plantea presentar esta estructura de informes con un ejemplo práctico haciendo uso de datos de un comercio electrónico. Con esto el asistente tendrá un alcance de las opciones que tiene un usuario para el manejo de datos de analítica digital y las diferentes librerías que pueden resultar útiles.

Palabras y frases clave: Análisis digital, automatización de informes, Google Analytics, shiny.

PLANIFICACIÓN ESTRATÉGICA CON R

Belén M Fernández de Castro[1], Teresa Veiga Rodríguez[1]

[1] DGA y PMO de ABANCA

RESUMEN

La Dirección General Adjunta (DGA) de Planificación Estratégica y PMO de ABANCA está formada por un equipo multidisciplinar que desempeña muy diversas tareas de relevancia para la Entidad. En el área de Planificación y Estudios dichas tareas son, por ejemplo, el desarrollo de los Planes Estratégicos, la presupuestación anual, dar respuesta a los distintos requerimientos del supervisor (Stress Test, ICAAP,...) o dar soporte a los órganos de decisión del Banco.

Muchas de las tareas citadas toman como punto de partida una hipótesis sobre la evolución a futuro de ciertos indicadores macroeconómicos. Además, otras áreas del banco, como por ejemplo el área de riesgos, necesita también una previsión de indicadores macroeconómicos que influyen en ciertos modelos internos de previsión, como puede ser el caso de la previsión de entrada en mora.

Para garantizar la coherencia de todos estos ejercicios, desde el área de Planificación Estratégica se han desarrollado modelos estadísticos que permiten generar un cuadro de indicadores macroeconómicos, con previsiones a 3 años, que sirven de soporte a toda la entidad. Estos modelos macroeconómicos se han desarrollado con el software estadístico R. Además, se ha implementado una aplicación que permite a los usuarios interactuar de manera sencilla permitiendo, por ejemplo, actualizar los datos, revisar las previsiones y los modelos estimados y obtener un cuadro de previsiones completo.

Palabras y frases clave: Web scraping, series temporales, modelización y creación de aplicaciones.

1. UNA SOLUCIÓN PARA GENERAR CUADROS MACORECONÓMICOS

Una de las principales tareas del área de Planificación y Estudios de la DGA de Planificación Estratégica y PMO de ABANCA tiene que ver con el seguimiento y previsión de indicadores macroeconómicos. Para facilitar esta tarea se han desarrollado diferentes utilidades en R haciendo uso de diversos paquetes y librerías, como pueden ser:

- Para el seguimiento macroeconómico se emplean rutinas que permiten obtener de manera automática los datos de los diferentes indicadores económicos que proporcionan varias instituciones públicas, como es el caso del Banco de España o el IGE. Esta descarga de información se produce cada vez que se ejecuta la aplicación. Además, los procesos de descarga se han automatizado de manera que, en función de la

frecuencia de publicación de cada caso, la descarga se produce sin la necesidad de ejecutar el código por parte del usuario.

- Gran parte de la carga de trabajo que se realiza con R está relacionada con el análisis exploratorio de datos y el estudio de indicadores, empleando tanto modelos para desestacionalizar las series temporales y poder trabajar con ellas corregidas, como modelos predictivos, entre los que destacan la metodología Box-Jenkins, modelos multivariantes y regresiones dinámicas. Actualmente se está trabajando también en la síntesis de distintas series macroeconómicas en un único indicador que dé cuenta del estado de la economía.

- A menudo surge la necesidad de plasmar los resultados en un documento, bien porque tienen que ser trasladados a un supervisor o a efectos de documentación interna. Esto se puede hacer en R a través de la librería Markdown, que permite generar documentos (doc, html o pdf) de manera simultánea a la ejecución del código, mostrando los resultados tanto en tablas, gráficos o mismo el código si se desea.

- Para dotar de herramientas a aquellos usuarios no conocedores del lenguaje R se ha desarrollado una aplicación con la librería Shiny, en la que se incluyen las funcionalidades anteriormente mencionadas. Las motivaciones y prestaciones por las que se ha optado por esta vía son, entre otras:

    - Facilita la visualización y el análisis de los datos y resultados de una manera más amena mediante las tablas y los gráficos dinámicos.
    - Permite a los usuarios realizar análisis de sensibilidad de las predicciones obtenidas con los modelos. Esta tarea consiste en estudiar cómo reaccionan las variables de estudio a diferentes escenarios.
    - A partir de un escenario base fijado, el usuario puede generar distintas versiones de escenarios adversos u optimistas a los que se les puede asignar una probabilidad de ocurrencia mediante un desarrollo interno.
    - En general, otorga cierta autonomía a los usuarios que no están habituados a trabajar con código R, haciéndoles partícipes de todo el proceso.
    - Garantiza la trazabilidad y la calidad de la información utilizada en las distintas áreas. Además, al tener un origen único se garantiza también la consistencia.
    - Actualmente se está trabajando en el "empaquetado" de estas aplicaciones, de manera que no sea necesario acceder a R para ejecutarlas. Con esto se facilitaría la tarea de compartirlas dentro de la Entidad.

Referencias

[1] Adrian A. Dragulescu and Cole Arendt (2018). xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.1. https://CRAN.R-project.org/package=xlsx.

[2] Bernhard Pfaff (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars. Journal of Statistical Software 27(4). URL http://www.jstatsoft.org/v27/i04/.

[3] Dan Vanderkam, JJ Allaire, Jonathan Owen, Daniel Gromer, Petr Shevtsov and Benoit Thieurmel (2017). dygraphs: Interface to Dygraphs Interactive Time Series Charting Library. R package version 1.1.1.4. https://CRAN.R-project.org/package=dygraphs.

[4] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng and Winston Chang (2018). rmarkdown: Dynamic Documents for R. R package version 1.10. https://CRAN.R-project.org/package=rmarkdown.

[5] Kung-Sik Chan and Brian Ripley (2012). TSA: Time Series Analysis. R package version 1.01. https://CRAN.R-project.org/package=TSA.

[6] Ruey S. Tsay (2015). MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 0.33. https://CRAN.R-project.org/package=MTS.

[7] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny.

[9] Yihui Xie (2018). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.4. https://CRAN.R-project.org/package=DT.

VISUALIZACIÓN DA INFORMACIÓN ESTATÍSTICA UTILIZANDO *R-SHINY* E *R-MARKDOWN*

Noa Veiguela Fernández[1], Esther López Vizcaíno[1] e Ana Andión Hermida[1]

[1] Instituto Galego de Estatística (IGE)

RESUMO

No Servizo de Difusión e Información do Instituto Galego de Estatística (IGE) temos encomendada a tarefa, como o seu nome indica, de dar a coñecer o traballo do Instituto. A principal ferramenta nesta labor de difusión constitúea a páxina web. Dende a súa creación, fai xa case dúas décadas, evolucionou considerablemente, adaptándose aos novos desenvolvementos tecnolóxicos que tiveron lugar no campo do deseño web. Nos últimos tempos, a esta tarefa continua de adaptación técnica, súmanse novos retos:

- Deseñar produtos adaptados ao público actual, que demanda unha web dinámica coa que poder interactuar
- A necesidade de automatizar procesos e tarefas repetitivas, como a elaboración periódica de informes
- Buscar programas que permitan acercarse ao deseño web aos estatísticos, sen necesidade de contar cun coñecemento profundo das linguaxes de programación HTML, CSS ou JavaScript

Neste relatorio presentamos como estamos afrontando estes novos retos no IGE. Neste Instituto estamos comprometidos co desenvolvemento do software libre R e recorremos a el sempre que podemos na difusión da información estatística. Nesta ocasión servímonos, de novo, de dúas das súas librarías: *R-Shiny* e *R-Markdown*. A primeira permite crear aplicacións web dinámicas nas que o usuario se comunica co sistema informativo que está detrás da web, obtendo dela o que precisa en cada momento. A segunda libraría permite producir informes en diversos formatos (html, pdf, word), reducindo enormemente o tempo de traballo adicado a esta tarefa. Unhas das vantaxes de ambas as dúas librarías é que, sen contar cun gran bagaxe no deseño web, se poden crear páxinas moi completas e vistosas.

Palabras e frases chave: web dinámica, automatización procesos, *R-Shiny*, *R-Markdown*

1. INTRODUCIÓN

O décimo-quinto principio do "Código de boas prácticas das estatísticas europeas", adoptado polo Comité do Sistema Estatístico Europeo, establece que as estatísticas europeas deben presentarse "de forma clara e comprensible" e difundirse "de forma adecuada e conveniente" [1]. O cumprimento deste principio avalíase por medio dunha serie de indicadores, un dos cales establece que os servizos de difusión deberán utilizar "unha tecnoloxía moderna de información e comunicación, métodos, plataformas e estándares de datos abertos" [2]. Baixo este principio subxace a necesidade de adaptar os contidos aos distintos tipos de usuarios que visitan as

páxinas web dos organismos estatísticos [3]. Na actualidade, esta é a ferramenta máis empregada para difundir os resultados estatísticos e así se contempla no principio décimo-quinto do Código. Por tanto, os servizos estatísticos deben manter ao día os seus portais, aplicando os últimos desenvolvementos tecnolóxicos no campo do deseño web e impedindo a súa obsolescencia.

Para poder adaptar os nosos contidos e modos de difusión aos distintos tipos de usuarios, primeiro debemos saber cales son as súas necesidades informativas. Podemos establecer dúas grandes categorías de "oíntes": o usuario especializado, con experiencia no manexo de datos estatísticos e na navegación por webs desta índole, e o usuario non especializado, pouco acostumado ao manexo e interpretación de datos estatísticos. Este segundo grupo de usuarios busca nas webs de estatística un coñecemento xenérico da realidade socioeconómica que o rodea e acceder a el de forma sinxela. Ademais, tanto un grupo, coma outro, demandan cada vez máis unha web interativa que lles ofreza ferramentas de consulta e visualización personalizadas.

Ao primeiro grupo de usuarios dirixiremos produtos complexos, como a consulta multi-tabla, que proporciona acceso ás bases de datos onde se almacena a información estatística para que o usuario constrúa a súa propia táboa de resultados. Trátase dun formato de difusión *aberto* no que a carga de interpretar os resultados estatísticos recae sobre o propio usuario. Non obstante, para o segundo grupo de "clientes" pode ser máis recomendable recorrer a un formato de difusión *pechado*, que ofreza os resultados estatísticos xa tratados, mediante compendios de datos tabulados, acompañados de representacións gráficas e de notas explicativas [4]. Unha terceira vía para difundir a estatística consiste en combinar as dúas anteriores en aplicacións web dinámicas que permitan ao usuario un certo grao de autonomía á hora de decidir o que se consulta, pero ofrezan o resultado da mesma xa disposto en forma de táboas, gráficos e mapas. Ademais, é recomendable que as tres vías permitan, en maior ou menor medida, a descarga da información en formatos para a súa almacenaxe e a súa impresión, como as follas de cálculo ou os arquivos pdf.

O principio décimo-quinto do Código formula unha serie de retos na difusión de estatísticas, difíciles de acadar para os organismos de carácter rexional, coma o Instituto Galego de Estatística (IGE), cuxos equipos de traballadores e traballadoras son moito máis modestos que os dos organismos nacionais e supranacionais. Elo obriga ao persoal destes servizos estatísticos a acometer múltiples funcións, entre elas o deseño web, máis preto da rama da informática que da estatística. Non obstante, a necesidade agudiza o enxeño e, lonxe de rexeitar o desafío, no IGE propuxémonos desenvolver a web moderna e dinámica que demanda a cidadanía galega.

Para acometer as tres vías de difusión mencionadas no parágrafo anterior co equipo de que dispomos, debemos...

- Recorrer a automatizar procesos e tarefas que se repiten con frecuencia, como a elaboración periódica de informes de resultados, de forma que se minimice o tempo e o persoal empregado na súa redacción.
- Buscar programas que permitan acercarse ao deseño web aos estatísticos, sen necesidade de contar cun coñecemento profundo das linguaxes de programación HTML, CSS ou Javascript.

E todo elo cun custo económico adicional nulo, polo que os programas que empreguemos para logralo deben ser gratuítos.

No IGE estamos comprometidos co desenvolvemento do software libre R e recorremos a el para desenvolver varias das ferramentas de difusión da nosa web. Este relatorio versa sobre dúas das súas librarías: *R-Shiny* e *R-Markdown*. A primeira permite crear as

aplicacións web dinámicas que mencionamos anteriormente. A segunda produce de forma simultánea informes en diversos formatos (html, pdf, word). Ambas librarías poden conectarse coas bases de datos que almacenan a información difundida, de forma que capturen directamente e en tempo real os datos almacenados cando o usuario accede á aplicación ou ao informe. Ademais, todo o que conteñen (táboas, gráficos, mapas e incluso texto) pode programarse de forma que amosen a información máis actual dunha estatística, reducindo o tempo que habería de dedicar unha persoa á actualización periódica de informes mensuais ou anuais.

2. APLICACIÓNS WEB DINÁMICAS CON *R-SHINY*

Neste apartado expóñense as diversas funcionalidades da libraría *R-Shiny*, servíndonos das aplicacións desenvolvidas para a web do IGE a modo de exemplo.

3. AUTOMATIZACIÓN DE INFORMES PERIÓDICOS CON *R-MARKDOWN*

O apartado 3 dedicouse ao desenvolvemento do paquete *R-Markdown*; de novo, exporanse exemplos de informes realizados con el e difundidos na web do IGE para exemplificar o seu funcionamento.

4. CONCLUSIÓNS

Ao longo do relatorio púxose de manifesto a necesidade de adaptar os contidos da web do IGE aos distintos tipos de usuarios que a consultan, así como os formatos de difusión. Defínense dous tipos de usuarios: especializados (con facilidades á hora de interpretar datos estatísticos e habilidades na navegación polas webs desta índole) e non especializados (cun coñecemento xeral da nosa rama de estudo e non dados ao tratamento de datos). Presentáronse dous formatos de difusión web adaptados ás necesidades de cada grupo: as aplicacións web e os informes de resultados.

En ambos casos recorreuse ao software de código aberto R na súa elaboración; en particular, á libraría *Shiny*, que permite crear aplicacións web dinámicas, e ao paquete *Markdown*, que facilita a elaboración de informes en múltiples formatos. Os motivos polos que nos decantamos por este programa para personalizar os formatos de publicación web foron varios:

- Permite deseñar produtos de difusión interactivos sen necesidade dunha gran bagaxe nas linguaxes de programación web HTML, CSS ou Javascript.
- Os produtos elaborados con este programa poden conectarse coas bases de datos que almacenan a información difundida na web, de forma que capturen directamente e en tempo real os datos almacenados a petición do usuario.
- Os contidos das aplicacións e informes web (tales como táboas, gráficos e incluso texto) poden programarse para que amosen a información máis actual dunha estatística, sen requirir a actualización manual por parte dun estatístico.
- Posibilita a descarga da información en formatos que permiten a súa almacenaxe e impresión (como pdf e word).
- É gratuíto e nútrese das achegas dos usuarios, polo que se trata dun programa vivo en continuo proceso de desenvolvemento e mellora.

Referencias

[1] Comité do Sistema Estatístico Europeo (2017). Código de boas prácticas das estatísticas europeas. En rede

https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7, 10.

[2] Na súa redacción orixinal, o Código establecía tamén como indicador para avaliar o cumprimento do principio 15 difundir "copia impresa tradicional" nos casos que requirisen este formato. No IGE cremos que ofrecer un arquivo descargable, susceptible de ser impreso, pode ser aconsellable nalgúns casos que se detallarán na ponencia, como na publicación de informes e resumos de resultados.

[3] En UNECE (2009) tamén se pon o énfase en que a mensaxe debe adaptarse ao público obxectivo: "a primeira decisión importante que debes tomar é elixir con precisión unha audiencia: ¿para quen estou escribindo? Sinxelamente, o público é quen manda. En xeral, o que o público quere é o que deberías ofrecerlle"; Comisión Económica para Europa das Nacións Unidas (UNECE) (2009). Como facer comprensibles os datos. Parte 2: unha guía para presentar estatísticas. En rede:

https://www.unece.org/index.php?id=17568&L=3, 1.

[4] Veiguela Fernández, N. (2016). R-Shiny: una herramienta para mejorar la difusión de las operaciones del Sistema de Cuentas Económicas de Galicia, ponencia presentada nas XIX Jornadas de Estadística de las Comunidades Autónomas, Madrid. En rede:

http://www.jecas.es/2016_Madrid/ponencias/H2.pdf, 2.

VISUALIZACIÓN INTERACTIVA DE DATOS DE SAÚDE. [CREACIÓN DE DASHBOARDS CON Shiny]

Miguel Ángel Rodríguez Muíños[1]

[1] Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia.

Palabras e frases chave: R, rstats, shiny, dashboard

RESUMO

A Dirección Xeral de Saúde Pública pon á disposición dos profesionais e da cidadanía en xeral un espazo web onde atopar información, documentación, datos, software, ... referentes a aspectos relacionados directa ou indirectamente coa protección, prevención e promoción da saúde da poboación.

https://dxsp.sergas.gal

Esta web ten unha distribución temática (estilos de vida saudables, enfermidades transmisibles, sanidade ambiental, conductas adictivas, indicadores de saúde...). Ún deses apartados é o de "Datos" (https://www.sergas.gal/Saude-publica?idcatgrupo=11035) no que se publica información (case sempre numérica) sobre distintos sistemas de información.

Por mor da arquitectura tecnolóxica dispoñible a nivel corporativo no que se refire á xestión de contidos na web, a información que se está a ofrecer ate o momento é totalmente estática e hai que elaborar, nalgúns casos, "visualizacións" predefinidas (para cada suposto, cada ano, cada organización xeográfica, ...), traballo que resulta moi costoso en tempo e recursos humanos. Además, é necesario refacelo cando dispoñemos dunha actualización dos datos (nalgúns casos, anualmente; noutros, con máis frecuencia).

Este escenario, a medida que crece o apartado de "Datos", resulta cada vez máis complexo de xestionar e require un maior esforzo en horas e persoas.

Esta dirección xeral, plantexou a necesidade de "migrar" a un sistema de visualización dinámica de datos que permita que sexa o propio usuario o que manexe eses datos e constrúa as peticións que estime oportunas. Este sistema, se corresponde cun **dashboard** (interface gráfica que permite a análise de datos de diferentes fontes).

Despois do estudo das distintas posibles solucións (BIRT, Tableau, Qlik, ...) optouse por implementar un dashboard open source, gratuito e pouco agresivo co sistema informático corporativo. A solución foi usar R e Shiny (baixo ubuntu server 18.04 LTS).

Agora mesmo estamos en fase de desenvolvemento do primeiro proxecto (Sistema de Información de Mortalidade por Cancro de Galicia -SIMCA-) e os resultado xa son visibles en:

http://saudepublica.melisa.gal/SIMCA.

Na Figura 1 pode verse unha captura da aplicación web.

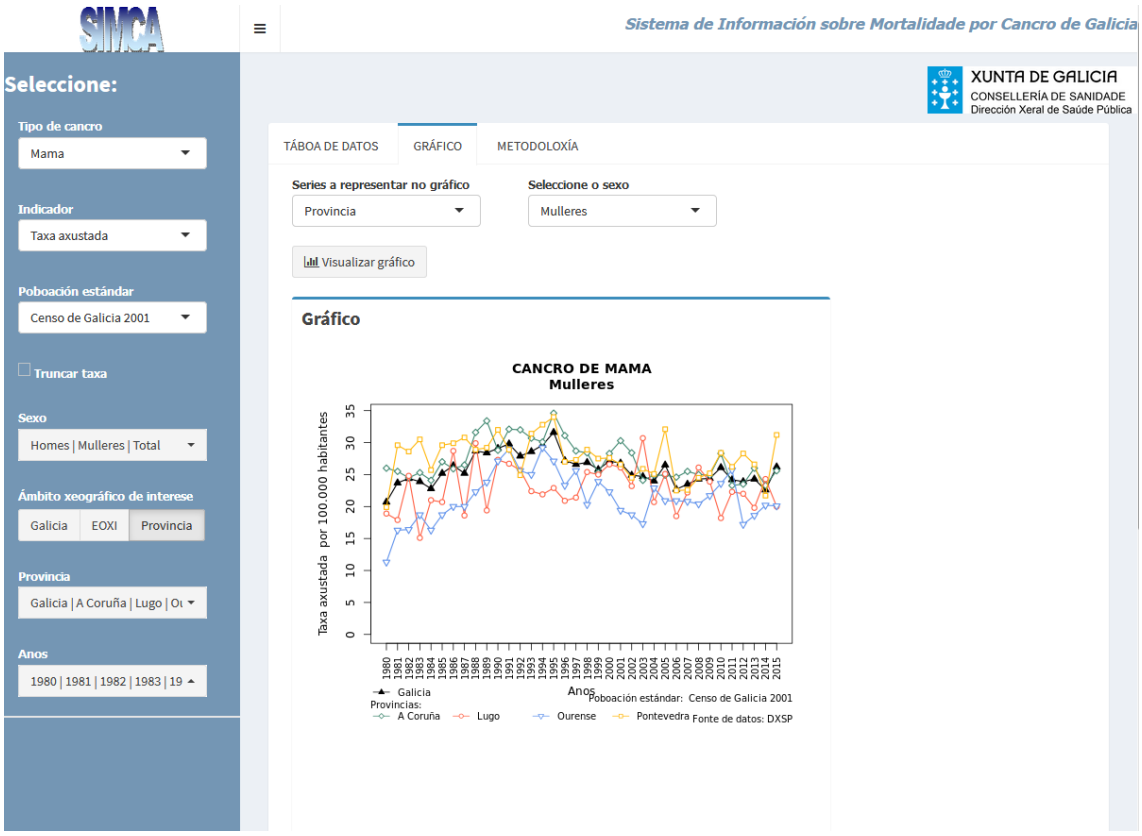


Figura 1: Dashboard do SIMCA

R EN LA ADMINISTRACIÓN PÚBLICA: XUNTA DE GALICIA. CÁLCULOS CON INTERVALOS DE FECHAS

Marcos Fernández Arias[1]

[1] Axencia para a Modernización Tecnolóxica de Galicia (AMTEGA), Xunta de Galicia

RESUMEN

Mediante R realizamos análisis cuantitativos para evaluar la prestación de diversos servicios. Algunos de estos servicios son prestados directamente por la administración y otros a través de empresas proveedoras.

En los contratos firmados, evaluamos el cumplimiento de los "acuerdos de niveles de servicio" (SLA) por parte de las empresas proveedoras.

Palabras y frases clave: service level agreement, regulatory compliance, OTRS, issue tracking systems, data intervals

1. INTRODUCCIÓN

Dentro del sector de servicios de tecnologías de la información, la Xunta de Galicia tiene firmados varios acuerdos con empresas proveedoras.
Algunos de estos contratos contienen cláusulas específicas de "acuerdo de nivel de servicio" (SLA) en donde se especifican plazos temporales admisibles y tiempos de atención y resolución. Se especifican también penalizaciones en caso de incumplimiento.

OTRS es un sistema open source de gestión de tickets con diversas prestaciones para gestionar llamadas y e-mails de clientes. Incluye gestión de tickets, flujos de trabajo y automatización aparte de la posibilidad de adaptaciones.
Se utiliza en gestión de servicios de tecnologías de la información para estructurar mejor las comunicaciones y tareas.

En la Axencia de Modernización Tecnolóxica de Galicia, OTRS es un sistema que resulta estructural e imprescindible para un funcionamiento interno eficiente, organizado y robusto.

2. ANÁLISIS DEL CUMPLIMIENTO DE SLA (SERVICE LEVEL AGREEMENTS)

En la Xunta de Galicia gestionamos mediante OTRS:

- un volumen de tickets con una ratio de creación de aproximadamente 15000 nuevos tickets cada mes
- y unos 30000 escalados mensuales ("escalado" es el envío de un ticket de un grupo a otro)
- gestionados por más de 1500 agentes (técnicos que actúan como intermediarios o realizan resolución de tickets)

- atendiendo a las necesidades de aproximadamente 23000 usuarios (registrados en Directorio Activo)

Utilizamos R y diversas librerías del ecosistema Tidyverse para evaluar los tiempos de atención y respuesta en los tickets, por parte de los distintos grupos.

Referencias

[1] OTRS AG. OTRS: leading service management suite. https://otrs.com/product-otrs/

THE IMPORTANCE OF TESTING R CODE

Nora M. Villanueva[1]

[1] Gradiant, Galician Research and Development Center in Advanced Telecommunications, Spain

ABSTRACT

In a software development framework, testing code is one of the good practices used to ensure that expected business systems and product features behave as expected.

Additionally, another case where software testing is incredibly important is when one provides code for other people to use. For example, if you have developed an R package that will be made publicly available, you should be sure that it works correctly, with quality and good performance. There are many tools available which provide utilities to help and develop software testing , particularly, `testthat` or `assertive` are some of them in the R context. In this talk, I will discuss how to incorporate testing in your work.

Keywords: Testing Code; R Packages; Test Driven Development; Software Development.

References

[1] Cotton, R.(2016). *assertive: Readable Check Functions to Ensure Code Integrity*, 2016. R package version 0.3-0.

[2] Cotton, R. (2017). *Testing R Code*. A Chapman and Hall.

[3] Wickham, H. (2011). testthat: Get Started with testing. *The R journal* 3, 5-10.

INTEGRACIÓN DE R EN QGIS

Ana Belén Buide Carballosa[1], María José Ginzo-Villamayor[2], Manuel Antonio Novo Pérez[1], Manuel Oviedo de la Fuente[1]

[1] Instituto Tecnolóxico de Matemática Industrial (ITMATI)
[2] Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela (USC)

RESUMO

QGIS permite a integración con paquetes estatísticos e sistemas de información xeográfica (SIX) de código aberto. Neste traballo, explicarase como configurar o sistema de procesamento de QGIS para executar un algoritmo implementado en R desde este.

Tamén se presentará un algoritmo para a obtención de rutas de escape para as brigadas que traballan na extinción dun incendio forestal, amosando a interfaz creada desde QGIS para a súa execución. Este algoritmo forma parte dun conxunto máis amplo de algoritmos desenvoltos dentro do proxecto Enjambre, que serven para a axuda na toma de decisións durante os incendios forestais.

Palabras e frases chave: QGIS, R, ruta de escape.

1. CONFIGURACIÓN DE R EN QGIS

QGIS, software libre e de código aberto, estase a converter nun sistema información xeográfica (SIX) líder do mercado, con módulos de xeoprocesamento similares ás ferramentas dispoñíbeis en SIX privativos, como ArcGIS. QGIS permite a integración con paquetes estatísticos e SIX, de código aberto. A continuación detállanse os pasos necesarios para executar un algoritmo de R en QGIS.
Para executar un algoritmo de R en QGIS o primeiro paso a levar a cabo é configurar R en QGIS e indicar a carpeta onde están localizados os arquivos binarios de R. Para isto, na interfaz de QGIS débese ir ao menú *Procesos* e logo seleccionar *Opciones*…. Na ventá de *Opciones de Procesos*, ver Figura 1, desprégase a pestana *Proveedores* e despois a de *R scripts*. Neste último despregable:

a) Marcase a opción *Activate*.
b) Na opción *Carpeta de R*, indícase o directorio no que está instalado no noso ordenador.
c) Na opción *Carpeta de biblioteca de usuario de R* indícase o directorio no que QGIS almacena as librerías de R.
d) Na opción *Carpeta de scripts de R* indícase o directorio no que QGIS gardará os scripts de R.
e) (Opcional) Actívase a opción *Use la versión de 64 bits* se o equipo de traballo ten 64 bits.
f) Clic sobre *Aceptar* para finalizar a configuración de R scripts en QGIS.

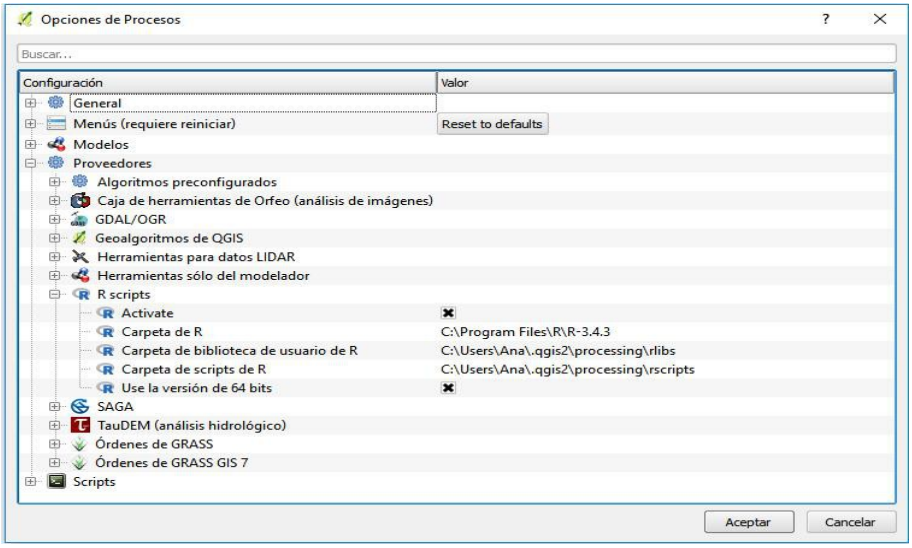


Figura 1: Opcións de configuración de R en QGIS.

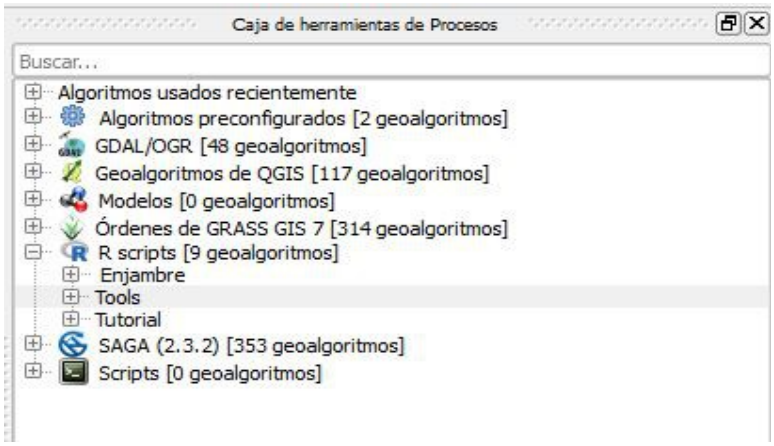


Figura 2: Caixa de ferramentas de procesado de QGIS.

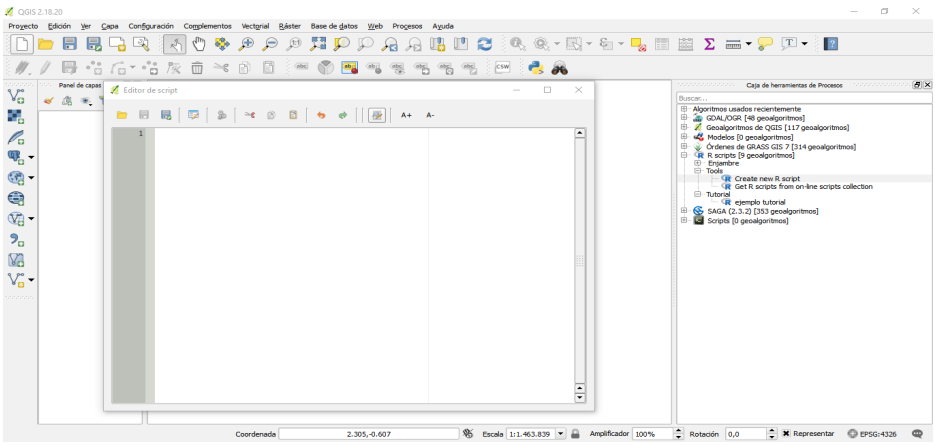


Figura 3: Creación dun script de R en QGIS.

Para abrir a caixa de ferramentas de procesos, no menú *Procesos* de QGIS clícase sobre *Caja de herramientas*. Deste xeito, esta abrirase no lado dereito da interfaz de QGIS, ver Figura 2. Feito isto, para crear un script de R en QGIS vaise ao despregable *R*

scripts e en *Tools* faise doble clic sobre a opción *Create new R script*. Deste xeito, ábrese unha nova ventá chamada *Editor de script*, ver Figura 3.

2. PROXECTO ENJAMBRE

O proxecto Enjambre, "misións críticas de emerxencias con medios aéreos tripulados e non tripulados en voo cooperativo", iniciado no 2015 e financiado polo CDTI, céntrase en tres prioridades da estratexia de seguridade nacional: ordenación de fluxos migratorios, protección ante emerxencias e catástrofes, e seguridade marítima, nas que o sector privado pode dar soporte ao sector público, aportanto novos servicios innovadores que faciliten a cooperación público-privada.
O obxectivo do proxecto é o desenvolvemento de tecnoloxías avanzadas para combatir os incendios forestais, entre elas a dun novo servicio de misións críticas de emerxencias onshore e offshore, que permita a actuación de forma cooperativa e segura de aeronaves tripuladas e non tripuladas nun mesmo espazo aéreo, dispoñendo de aeronaves non tripuladas que desempeñen exclusivamente tarefas de observación ao servizo de aeronaves tripuladas de intervención, facilitando a toma de decisións durante as operacións.
Desde ITMATI desenvólvense algoritmos para este proxecto, capaces de delimitar o perímetro dun incendio, ofrecer datos da súa temperatura, evitar colisións entre aeronaves, calcular rutas de escape para unha brigada que traballa na extinción, etc..

3. ALGORITMO RUTAS DE ESCAPE

Na extinción dun incendio forestal o uso de brigadas é un elemento esencial para o control do incendio desde a terra. Por isto, é necesario manter unha boa comunicación e organización das mesmas para que se enfrenten con seguridade a este, evitando que o incendio na súa evolución chegue a cercalas. Esta situación pode estar motivada porque o terreo situado por detrás das brigadas se vai secando polo efecto da enerxía que desprende o incendio, de maneira que aumenta a probabilidade de que se incendie de forma que a brigada quede atrapada, privándolle dunha ruta de escape.
O obxectivo do algoritmo que amosamos é calcular unha ruta de escape para a evacuación dos medios de extinción terrestres, proporcionándolle ao director de estinción unha ruta de escape segura para retirarse desde o incendio ata o destino seleccionado. Para isto empréganse funcións de tempo mínimo en base ás distintas velocidades de desplazamento dunha persoa en función da pendente do terreo, do tipo de vexetación e do tipo de vía. As rutas obtidas conectan a posición das brigadas de extinción coa zona de seguridade máis próxima, evitando os posibles obstáculos que van encontrar no camiño e non poden sortear.
Este algoritmo implementado en R, integrouse en QGIS coa elaboración dun R script, proporcionándolle ao usuario unha interfaz amigable para a execución do mesmo, ver Figura 4.
Para a obtención da ruta de escape para a brigada desde a posición que se atopa ata unha zona de evacuación segura dada de antemán faise uso dunha capa ráster, onde o valor do píxel indica o coste de desprazamento (en km/h) e dunha capa vectorial que define o perímetro do incendio. A capa ráster cos costes de desprazamento pode obterse a partir do modelo dixital do terreo, da capa de vexetación clasificada en modelos de combustible e das capas vectoriais correspondentes a vías e obstáculos da zona.
Os datos de entrada do algoritmo son principalmente obxectos espaciais, datos vectoriais e ráster, cuxo tratamento se levou a cabo empregando os paquetes `sp` e `raster` de R, que dispoñen dunha amplia variedade de funcións para o análisis

espacial. Nos cálculos realizados foi preciso rasterizar capas vectoriais, un proceso que se pode levar a cabo empregando a función `rasterize()` do paquete `raster`, pero que dependendo das características do noso ordenador pode tardar máis ou menos tempo. Por isto último e dado que a execución do algoritmo debía ser o máis rápida posible, recurrimos a función `gdal\_rasterize()` do paquete `gdalUtils`, que permite optimizar o proceso de rasterizar capas vectoriais. Para o cálculo de distancias e rutas empregáronse funcións do paquete `gdistance`.

A execución do algoritmo desde a ventá implementada, Figura 4, finaliza coa obtención dunha ventá en formato HTML coa salida de R e coa carga automática dos resultados na vista de mapas de QGIS, Figura 5.

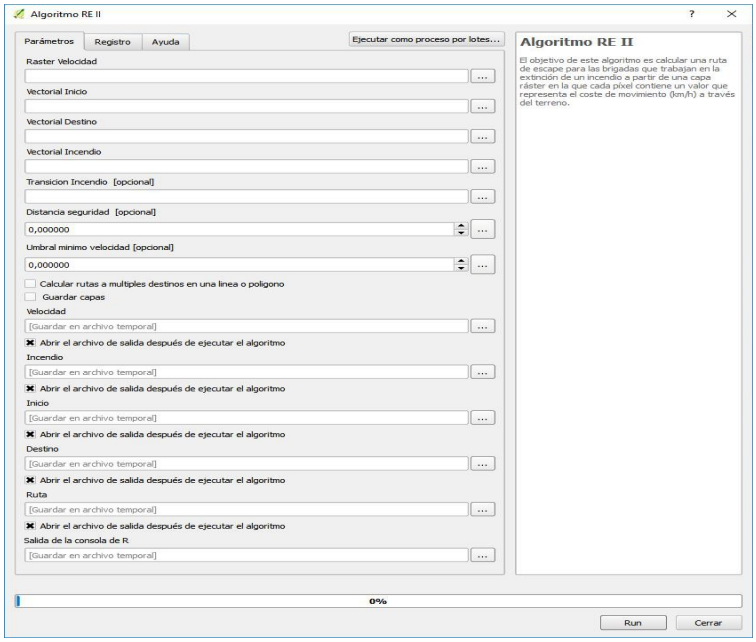


Figura 4: Interfaz do algoritmo creada a partir dun script de R en QGIS.

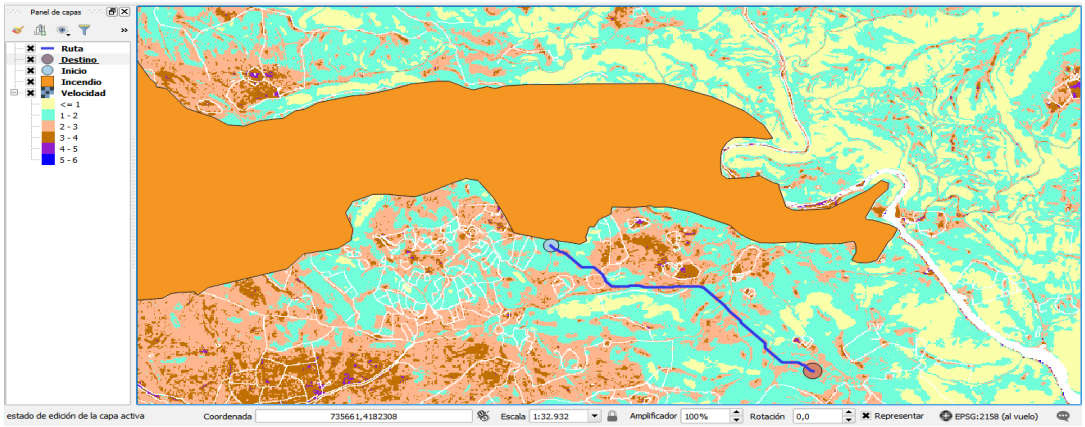


Figura 5: Resultado da execución do algoritmo. En cor laranxa represéntase o incendio, o punto azul claro fai referencia a posición da brigada e o marrón a zona de evacuación segura. En cor azul escuro amósase a ruta de escape.

AGRADECEMENTOS

Los investigadores Ana Buide, Manuel Novo, Manuel Oviedo y Mª José Ginzo agradecen el apoyo al proyecto ENJAMBRE del CDTI.

ALGORITMO DE EFICIENCIA EN LAS DESCARGAS DE AGUA

Manuel Antonio Novo Pérez [1], Ana Belén Buide Carballosa[1] e Mª José Ginzo Villamayor[2]

[1] Instituto Tecnológico de Matemática Industrial (ITMATI)
[2] Departamento de Estadística, Análisis Matemático y Optimización. Universidade de Santiago de Compostela

RESUMEN

El proyecto ENJAMBRE tiene como objetivo el desarrollo de un servicio de misiones críticas de emergencias onshore y offshore, que permita la actuació de forma cooperativa y segura de aeronaves tripuladas y no tripuladas en un mismo espacio aéreo, disponiendo de aeronaves no tripuladas que desempeñen exclusivamente labores de observación al servicio de las aeronaves tripuladas de intervención, facilitando la toma de decisiones durante las operaciones.

Dentro de este proyecto se ha desarrollado, entre otros, un algoritmo llamado Algoritmo de Eficiencia en las Descargas de Agua, que tiene como objetivo medir la eficiencia de las descargas de agua realizadas por los medios aéreos que trabajan en la extinción de un incendio forestal, entre dos instantes de tiempo. Para ello, el algoritmo utiliza las monitorizaciones de las aeronaves y las imágenes térmicas del incendio para generar capas ráster con la huella de agua estimada de los medios aéreos y una capa ráster en la que se mide la eficiencia de las descargas realizadas por los medios aereos.

El algoritmo se ha programado en R, en concreto, se ha utilizado el paquete `raster` ([1]), que permite leer imágenes ráster, modificarlas y crear otras nuevas. Esto permite al algoritmo manejar datos de entrada con ese tipo de información, así como también le permite generar imágenes ráster. En la Figura 1 se puede ver la huella de agua estimada generada por el algoritmo de las aeronaves que participaron en un incendio real.

Palabras y frases clave: Descargas de agua, eficiencia de descargas, ENJAMBRE, ráster.

1. METODOLOGÍA

La estimación de la huella de agua se ha basado en las fórmulas utilizadas en [2]. De esta forma se supone que las descargas son elípticas y se calculan, utilizando los modelos de regresión de la referencia, la longitud

$$L = -39{,}8116 + 0{,}0327586Q + 0{,}3333H + 0{,}556213V$$

y la anchura

$$W = 6{,}19933 - 0{,}0792313L + 0{,}00466447Q + 0{,}137522H - 0{,}0152993V$$

a partir del volumen de la descarga de agua (Q), la altura de la misma (H) y la velocidad de la aeronave (V), orientando la elipse seguún la dirección de la aeronave en el momento de la descarga.
En base a las fórmulas anteriores, el algoritmo programado en R, utilizando el paquete `raster`, sigue los siguientes pasos:

1. Definición de los datos de entrada. En este caso, dos imágenes térmicas consecutivas del incendio y las monitorizaciones de las aeronaves que participan en él.
2. Se cambia la extensión y resolución de las imágenes térmicas a una común para poder compararlas.
3. Se obtiene el ráster de diferencias de temperaturas.
4. Se obtiene el ráster de la huella total de las descargas para cada aeronave en base a sus monitorizaciones y a las fórmulas presentadas anteriormente, suponiendo que las descargas de agua son homogéneas.
5. Se obtiene la huella de agua total, resultante de la suma de las capas obtenidas en el punto anterior.
6. Se obtiene el ráster con los indicadores de la eficiencia de las descargas, donde cada píxel contiene la disminuci´on de temperatura entre los litros descargados, calculado a partir de los ráster de los pasos 2 y 5.
7. El algoritmo también puede funcionar en caso de carecer de imágenes térmicas, algo común si se utiliza en tiempo real. En dicho caso obtiene solamente la última descarga realizada por cada una de las aeronaves.

2. RESULTADOS

El algoritmo devuelve una capa ráster donde el valor de cada píxel se corresponde con los litros de agua descargados por la aeronave en ese punto, así en la Figura 1 se puede ver la huella de agua estimada generada por el algoritmo de las aeronaves que participaron en un incendio real y para cada aeronave que ha participado en la extinción del incendio, devuelve el número de descargas que ha realizado entre los dos instantes de tiempo que se están considerando.

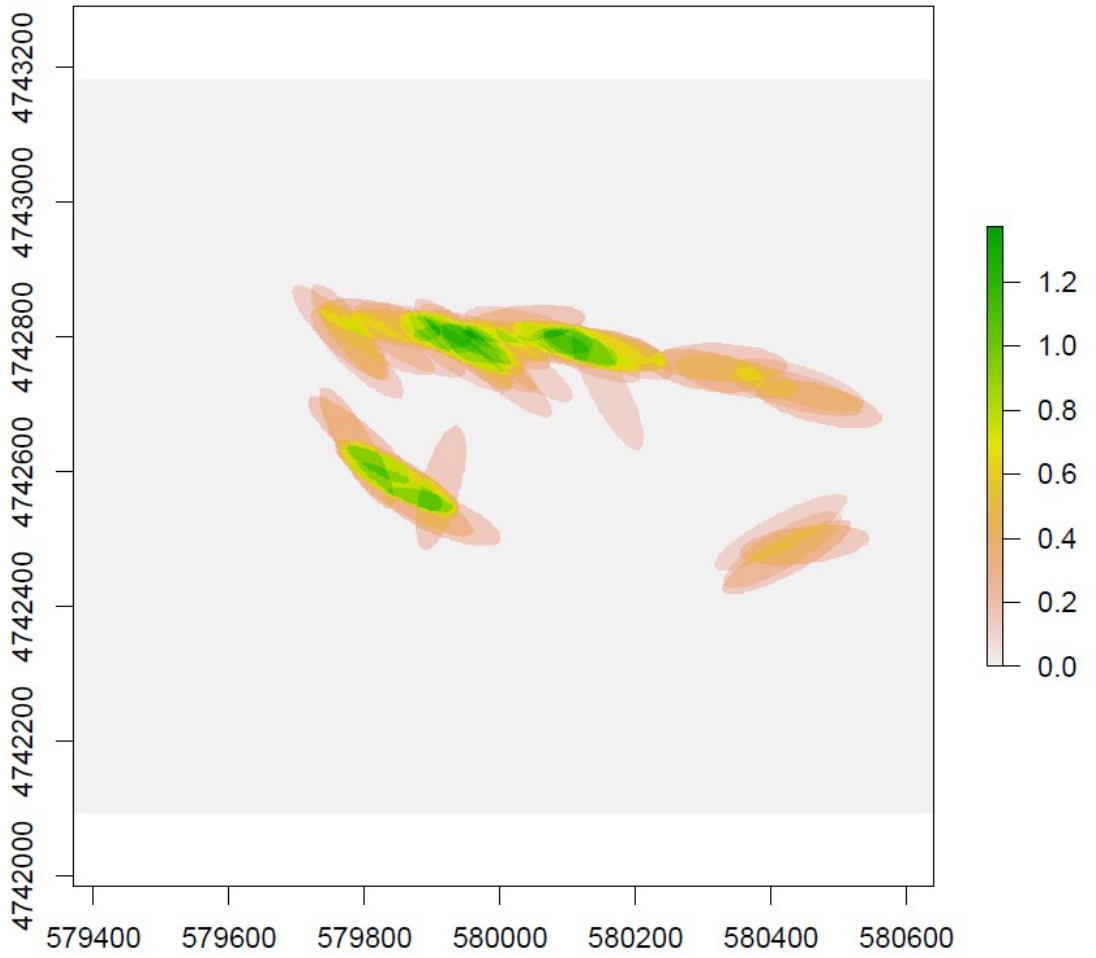


Ilustración 1: Representación en R de la capa ráster con la huella de agua estimada por el algoritmo de las aeronaves que participaron en un incendio. El valor de cada píxel viene dado por la cantidad de agua descargada (en litros) en el mismo.

AGRADECIMIENTOS

Los investigadores Manuel Novo, Ana Buide y María José Ginzo agradecen el apoyo del proyecto ENJAMBRE del CDTI.

Referencias

[1] Robert J. Hijmans (2017). raster: Geographic Data Analysis and Modeling. R package version 2.6-7.
URL https://CRAN.R-project.org/package=raster.
[2] Rodríguez y Silva F., González-Cabán A. (2010), "SINAMI": a tool for the economic evaluation of forest fire management programs in Mediterranean ecosystems. *Int. J. Wildl. Fire*, 19, pp. 927-936.

R Y HPC (USO DE R EN EL CESGA)

Aurelio Rodríguez[1]

[1] Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia (CESGA)

RESUMEN

El uso de R en un centro de HPC hoy en día permite reducir los tiempos de ejecución notablemente así como abordar problemas de mayor dimensión. Así mismo nos proporciona un lenguaje donde es posible implementar usos complejos de las infraestructuras HPC facilitando su uso.

Palabras y frases clave: R, CESGA, HPC.

1. INTRODUCCIÓN

Un centro HPC como el CESGA permite abordar problemas computacionales de dimensiones mayores mediante el uso de ordenadores muy potentes y los paradigmas de paralelización. Sin embargo presenta varias barreras que el usuario debe afrontar en su utilización: uso de un sistema operativo diferente (UNIX/LINUX), un sistema de colas, sistemas de ficheros compartidos (en red), etc. Desde el CESGA se intenta minimizar estas barreras con varias estrategias. Una de ellas es aprovechar APIs/lenguajes cercanos a un gran grupo de usuarios como es R para proporcionarles un lenguaje común que les facilite el uso de una infraestructura compleja de HPC.

2. USO DE UN SUPERCOMPUTADOR (FinisTerrae)

La arquitectura actual más habitual de un supercomputador es la arquitectura cluster donde un supercomputador está formado por numerosos nodos con distintas funcionalidades trabajando coordinadamente usando una o varias redes de interconexión. Para su uso el usuario debe ser capaz de conectarse remotamente al superordenador habitualmente con distintos protocolos (SSH es el más común). Esta primera conexión se establece con los login nodos donde el usuario obtiene una sesión limitada tanto en recursos como en tiempo pero que le va a permitir definir y enviar a ejecución el trabajo demandante a ejecutar.

En el siguiente esquema se resume la forma habitual de uso del FinisTerrae [1]:

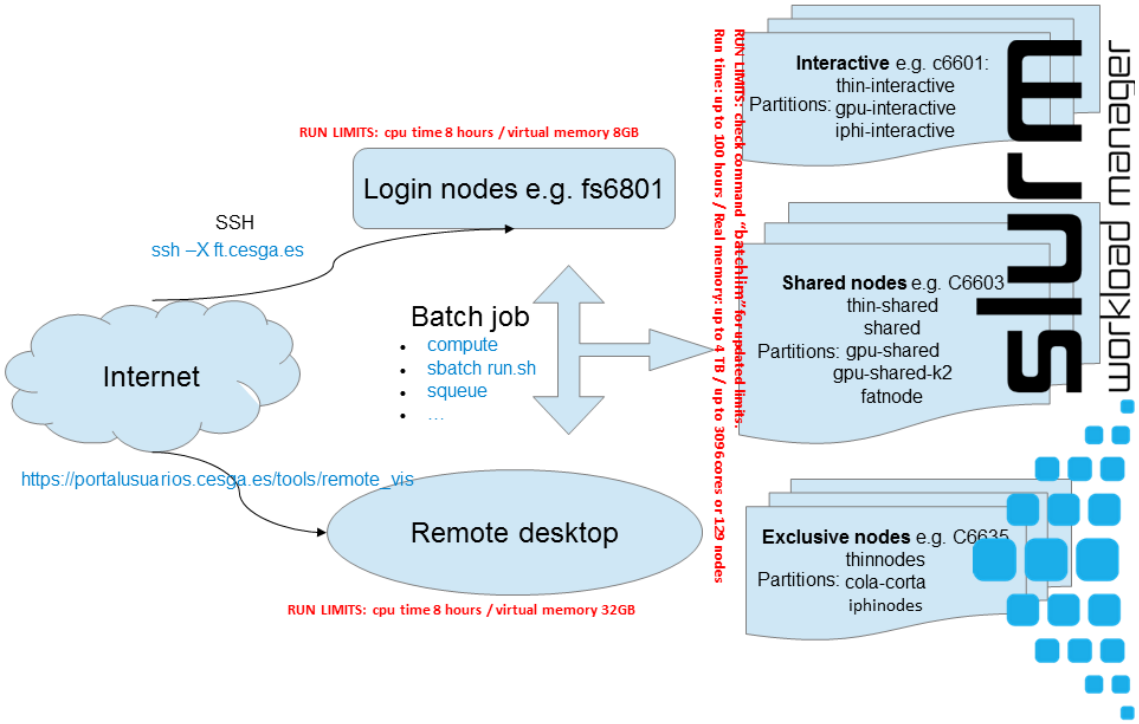


Figura 1: Uso FinisTerrae

3. R EN EL FinisTerrae

Desde el CESGA se proporciona las ultimas versiones de R preinstaladas con las siguientes características:

- Disponibilidad de interfaces gráficos para su uso como pueda ser Rstudio. Aunque el soporte de la versión server de Rstudio no es viable para el CESGA por motivos de seguridad, el acceso a Rstudio mediante un escritorio remoto es totalmente viable.
- El usuario puede instalar los paquetes que necesite de forma transparente en su cuenta e incluso mantener diferentes combinaciones de paquetes mediante la configuración adecuada del entorno (variable R_LIBS_USER).
- Compiladas usando las últimas librerías matemáticas de algebra líneal proporcionadas por Intel (Intel MKL) de manera que el rendimiento sea el óptimo y a su vez sea posible una paralelización automática en memoria compartida para este tipo de operaciones.
- Paquetes espécificos de HPC[2]: uso del sistema de colas y paralelización

Con estas características se posibilita que el usuario exclusivamente usando un interfaz gráfico de R como Rstudio pueda usar el FinisTerrae de forma completa.

4. CONCLUSIONES

Desde el CESGA se intenta facilitar el uso de R en un entorno de HPC proporcionando versiones optimizadas, adaptables por el usuario a sus necesidades y con los paquetes específicos que permiten el uso de un entorno HPC desde el lenguaje R.

Referencias

[1] Guía de uso del FinisTerrae. Disponible en: https://www.cesga.es/en/paginas/descargaDocumento/id/210

[2] CRAN Task View: High-Performance and Parallel Computing with R. Disponible en: *https://cran.r-project.org/web/views/HighPerformanceComputing.html*.

NOVAS LIBRERÍAS PARA O CONTROL ESTATÍSTICO DA CALIDADE (qcr) E ESTUDOS INTERLABORATORIO (ILS) NO CONTORNO DA INDUSTRIA 4.0

Salvador Naya[1], Javier Tarrío-Saavedra[1], Rubén Fernández-Casal[1] e Miguel Flores[2]

[1] Departamento de Matemáaticas. Grupos MODES, CITIC e ITMATI. Universidade da Coruña.

[2] Escuela Politécnica Nacional. Quito, Ecuador.

ABSTRACT

In this work, the new versions of the qcr (quality control review) and the ILS (Inter Laboratory Study) will be presented. These are two R packages that allow the study of control charts with functional data, the qcr library, and consistency between different laboratories, the ILS library. Examples of application are proposed for framed data within the so-called intelligent industry or Industry 4.0.

Palabras e frases chave: Control de Calidade, Consistencia, Estudos Interlaboratorio, Industria 4.0.

1. INTRODUCIÓN

O concepto da Industria 4.0 ou cuarta revolución industrial, representa unha nova era para a control de procesos. Un dos seus obxectivos é a implantación da chamada "fábrica intelixente", capaz de maior adaptabilidade ás necesidades dos procesos, así como a unha asignación de máis eficiente dos recursos. Neste contexto global da Industria 4.0, entendida como toda acción orientadaerimentos que permiten comparar laboratorios pero tamén distintas máquinas ou instrumentos de medición.

Por outra parte, o empleo de sensores e instrumentos de medida, cada vez m#áis sofisticados e capaces de proporcionar unha gran cantidade de datos dun n#úmero cada vez maior de variables críticas presentan novos retos para o control da calidade. Este tipo de problemas están enmarcados dentro dun campo importante da Industria 4.0, o do Big Data. Ademais o habitual é que este tipo de datos procedentes de sensores sexan datos funcionais (FDA). Para esta nova situación son precisas ferramentas para o seu debido tratamento estatístico, como as duas librerías de R que aquí se presentan (Naya, 2017).

A librería `qcr`, permite facer todos os estudos de control estatístico da calidade ampliando tamén o espectro a gráficos de control de tipo non paraméetrico e para datos funcionais. A aplicación propoñerase a problemas de eficiencia enerxética, onde se busca detectar anomalías mediante o emprego de detección de atípicos.

No caso da librería `ILS` permite o estudo completo de experimentos entre varios laboratorios para estimar a consistencia. Ademais esta librería pode tamén aplicarse noutros contextos para detección de atípicos en exemplos ligados a industria 4.0, como son aqueles vinculados a datos obtidos de sensores.

Estas das novas librerías aportan ferramentas para o tratamento de datos de tipo funcional, que ademas poden ser usadas sobre o mismo conxunto de datos, xa que unha está pensada para o control de calidade e a outra para diseñar experimentos

que permiten comparar laboratorios pero tamén distintas máquinas ou instrumentos de medición.

2. LIBRERÍA qcr

O paquete qcr inclúe un conxunto completo de ferramentas de control de calidade estatística (SQC), ferramentas univariantes e multivariantes que completan e aumentan as técnicas SQC dispoñibles en R.
Combina procedementos flexibles, tradicionais e novos SQC para abordar problemas reais de control de calidade na industria e o consultor. Ademais de integrar diferentes paquetes en R dedicados a SQC (qcc, MSQC), proporciona novas ferramentas paramétricas non-paramétricas moi útiles cando a suposición gaussiana non se cumpre.
Este paquete proporciona o conxunto máis completo de funcións en R para calcular os gráficos de control de atributos e variables, dende un punto de vista paramétrico e non paramétrico, dun xeito univariante ou multivariante. Para ser aplicado en problemas reais da industria, e especialmente de Indsutria 4.0, permite estimar os límites de control e monitorizar as variables críticas dunha forma máis automática e práctica.
O paquete proposto permite estimar os límites de control univariantes e facer gráficos de tipo estándar como os de Shewhart e tamen outros menos frecuentes como os EWMA e CUSUM. Ademais inclúe funcións para realizar gráficos de control multivariante como o da T2 de Hotteling, MEWMA e MCUSUM.
Ademais, permite usar novas alternativas non paramétricas baseadas na profundidade de datos como os champados gráficos r e Q e de novos gráficos con datos funcionais.

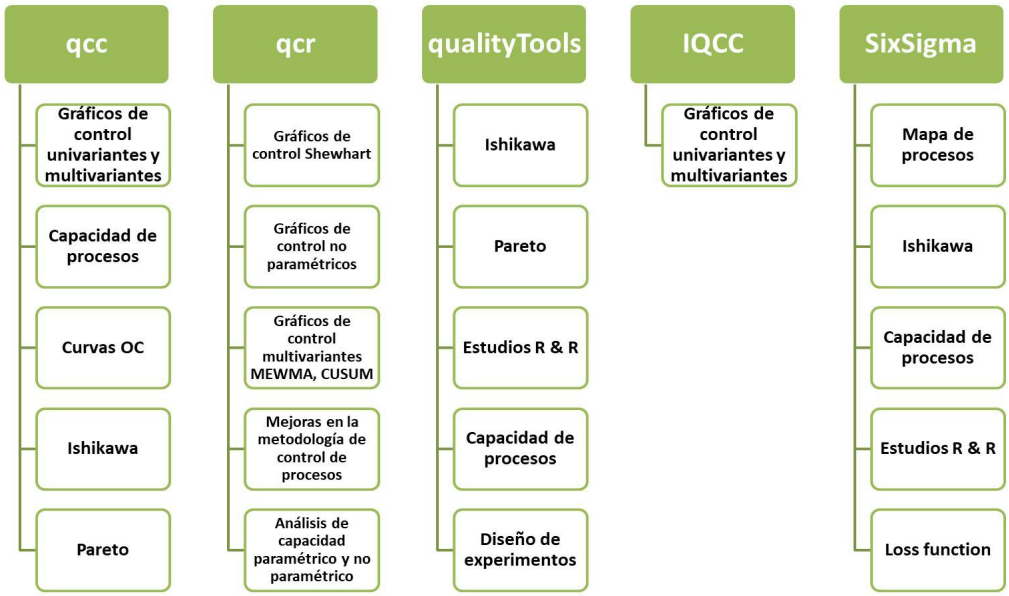


Figura 1: Paquetes de R para control de calidade e as metodologias incluidas.

3. LIBRERÍA ILS

As ferramentas estatísticas que facilita o paqute ILS permiten facer estudos interlaboratorio, tanto coa visión univariada, con base na norma ASTM E691 e normas ISO5725, como para casos de tipo funcional.

Entre aa opcións están o cálculo dos índices de Mandel para identificar os laboratorios que proporcionan resultados cuantivamente diferentes entre sí e estimar a consistencia para diseños entre varios laboratorios. Ademais, permite probar a presenza de valores atípicos a través das probas de Cochran e Grubbs.
Outra opción deste paquete é facer un Análisis de Varianza (ANOVA), incluíndo a prova F e a proba de Tukey para probar as diferenzas entre as medias da variable correspondentes aos distintos laboratorios.
Unha das novidades desta librería é incorporar ferramentas para levar a cabo unha proba ILS a partir de datos funcionais. Polo tanto, este paquete permite ter en conta a natureza funcional dos datos obtidos polas técnicas experimentais correspondentes á sensorización tan frecuente na Industria 4.0.
O paquete ILS permite estimar as estatísticas funcionais, H (t) e K (t) propostas por primeira vez no artigo de Flores et al. (2018a) e Flores et al. (2018b).

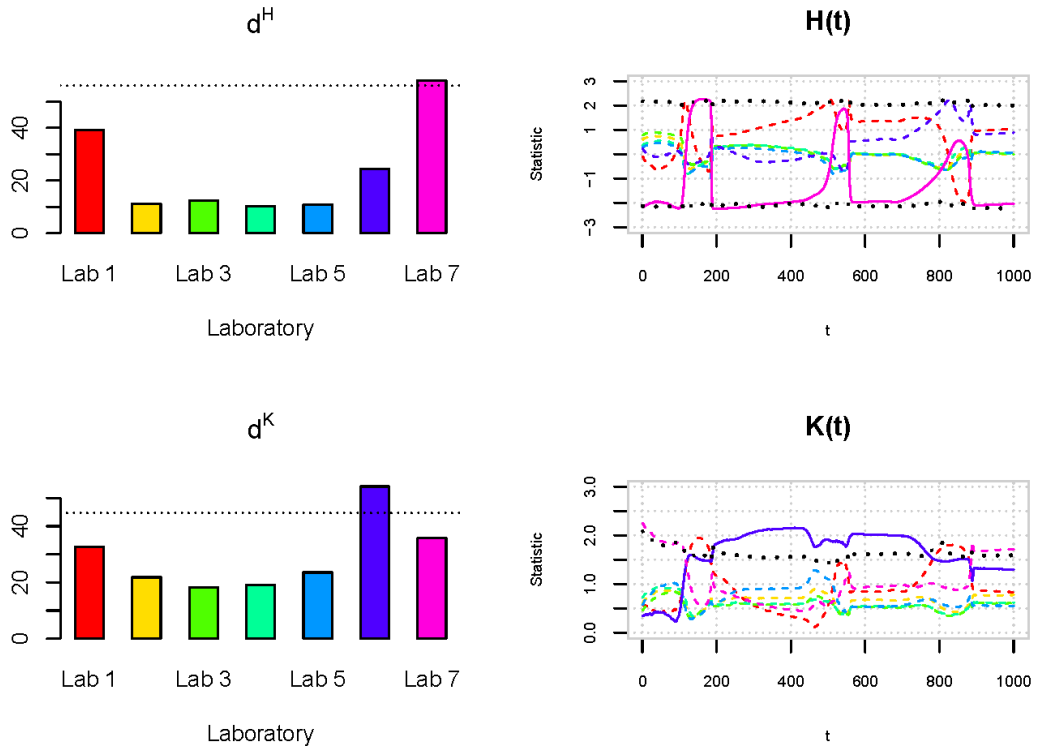


Figura 2: Gráfico estudio ILS con datos funcionales de 7 laboratorios de curvas TGA.

3. CONCLUSIÓNS

O desenvolvemento de novas metodoloxías, neste caso FDA, son de vital importancia para a xestión dos novos paradigmas e tipos de datos na dixitalización da industria. O seu desenvolvemento e aplicación é absolutamente necesario, non só para garantir a calidade dos produtos, servizos, procedementos, instrumentos e laboratorios, senón tamén para aumentar a calidade dos procesos de medición, reducir os custos de estudos incompletos e mellorar a análise dos resultados. En definitiva, facer unha mellora na calidade final do producto ou servicio.
Estes novos paquetes aquí presentados, o ILS e o qcr, esperamos, que máis pronto que tarde, sexan de uso común polos investigadores e responsables de laboratorio e empresas. A súa aplicación será de gran importancia no novo contexto da Industria

4.0, donde os métodos estatísticos xa teñen unha gran presencia e se prevé que tendrán un maior futuro.

AGRADECEMENTOS

Salvador Naya, Javier Tarrío-Saavedra e Rubén Fernández-Casal agradecen o proxecto MTM2017-82724-R (MINECO) e Miguel Flores o proxecto PII-DM-002-2016 da Escuela Politécnica Nacional de Ecuador.

Referencias

[1] Flores, M., Naya, S., Tarrío-Saavedra, J., Fernández-Casal, R. (2017). *Functional data analysis approach of Mandelâsh and k statistics in Interlaboratory Studies. In Functional Statistics and Related Fields*. A Coruña. Springer.

[2] Flores, M., Tarrío-Saavedra, J., Fernández-Casal, R. y Naya, S. (2018). Functional extensions of Mandel's h and k statistics for outlier detection in interlaboratory studies. *Chemometrics and Intelligent Laboratory Systems*, 176, 134-148.

[3] Flores, M., Tarrío-Saavedra, J., Fernández-Casal, R., Bossano, R., Naya, S. (2018). ILS: An R package for statistical analysis in Interlaboratory Studies. *Chemometrics and Intelligent Laboratory Systems*, 181, 11-20.

[4] Naya, S. (2017). *Industry 4.0. An Opportunity for the Relationship Between University and Shipbuilding in the Future*. Proceedings of the 25th Pan-American Conference of Naval Engineering. Panamá. Springer.

bookdown: UN PAQUETE DE R PARA A CREACIÓN DE LIBROS

Rubén Fernández-Casal[1], Tomás R. Cotos-Yáñez[2]

[1] Departamento de Matemáticas, Universidade da Coruña, 15071, A Coruña, España

[2] Departmento de Estatística e I.O., Universidade de Vigo, 32004 Ourense, España

RESUMO

O paquete bookdown de R permite escribir libros empregando R-Markdown de forma sinxela. Seguindo a filosofía de Markdown, podense crear libros en distintos formatos (HTML / PDF / Word / epub/ ...). Ademáis de permitir empregar as extensións Markdown de Pandoc (notas ao pé de páxina, táboas, citas, ecuacións LaTeX, ...), pódense empregar extensiones Markdown para libros (lendas de figuras e táboas, numeración e referencias cruzadas de figuras / táboas / seccións / ecuacións / teoremas / exemplos / ..., widgets HTML, ...).

Palabras e frases chave: Libros, R, Markdown, bookdown, Pandoc.

1. INTRODUCIÓN

O paquete bookdown [1] de R permite escribir libros de forma sinxela. Sen preocuparse moito polos detalles pódense crear libros en distintos formatos (HTML / PDF / Word / epub/ ...). Ademáis de permitir empregar as extensións Markdown de Pandoc (notas ao pé de páxina, táboas, citas, ecuacións LaTeX, ...), pódense empregar extensiones Markdown para libros (lendas de figuras e táboas, numeración e referencias cruzadas de figuras / táboas / seccións / ecuacións / teoremas / exemplos / ..., widgets HTML, ...).

O libro xerarase (especificados os formatos de saída) a partir dunha serie de documentos de R Markdown [2], cada un correspondente a cada capítulo. Na web https://bookdonw.org aparecen exemplos de libros creados por bookdown. Altamente recomendable é o recente libro de Yihui Xie *bookdown: Authoring Books and Technical Documents with R Markdown* [3].

2. REQUISITOS

1. Descargar e instalar RStudio IDE (versión superior a 1.0.0)
2. Instalar o paquete de R bookdown
3. Se se desexa obter o libro en formato laTeX/pdf precísase un compilador de laTeX.

Obviamente, un requisito é ter instalado R. Cando se instala o paquete bookdown, instálanse automáticamente os paquetes `knitr` e `Rmarkdown`. Un documento R Markdown (*.Rmd) compílase primeiro a Markdown (*.rd) usando o paquete `knitr` e con posterioridade Markdown o compila (por exemplo a LaTeX ou HTML) usando Pandoc. Pandoc non é un paquete de R, pero ven incluido no RStudio.

3. ESTRUTURA BÁSICA

Aínda que non é necesario o Rstudio é recomendable para principiantes. Ademáis ven cun *bookdown-demo* que pode ser usado de modelo inicial. A estrutura básica ven dada polos seguintes arquivos do directorio do proxecto:

- *index.rmd:* indícase por exemplo o título, autor, data do libro... ademáis de algunhas outras configuracións

```
---
title: "Authoring A Book with R Markdown"
author: "Yihui Xie"
date: "`r Sys.Date()`"
site: "bookdown::bookdown_site"
output:
bookdown::gitbook: default
documentclass: book
bibliography: ["book.bib", "packages.bib"]
biblio-style: apalike
link-citations: yes
---
```

- *output.yml*: Opcións do formato de saída de configuración *YAML.*

```
bookdown::gitbook:
  css: style.css
  config:
    toc:
      before: |
        <li><a href="./">A Minimal Book Example</a></li>
      after: |
        <li><a            href="https://github.com/rstudio/bookdown"
target="blank">
                                              Published             with
bookdown</a></li>
    edit: https://github.com/rstudio/bookdown-demo/edit/master/%s
    download: ["pdf", "epub"]
bookdown::pdf_book:
  includes:
    in_header: preamble.tex
  latex_engine: xelatex
  citation_package: natbib
  keep_tex: yes
bookdown::epub_book: default
```

- *bookdown.yml:* Arquivo de configuración onde se poden especificar parámetros opcionáis para compilar o libro.

```
book_filename: "SimulationR"
chapter_name: "Capítulo "
```

- *style.css* e *preamble.tex:* especifícanse as opcións de estilo e apariencia das saídas dos documentos en formato HTML e LaTeX, respectivamente.

```
style.css
--------
p.caption {
  color: #777;
```

```
  margin-top: 10px;
}
p code {
  white-space: inherit;
}
pre {
  word-break: normal;
  word-wrap: normal;
}
pre code {
  white-space: inherit;
}
preamble.tex
------------
\usepackage{booktabs}
```

- *book.bib:* arquivo onde se inclue as referencias bibliográficas.

```
@Book{xie2015,
  title = {Dynamic Documents with {R} and knitr},
  author = {Yihui Xie},
  publisher = {Chapman and Hall/CRC},
  address = {Boca Raton, Florida},
  year = {2015},
  edition = {2nd},
  note = {ISBN 978-1498716963},
  url = {http://yihui.name/knitr/},
}
```

4. PUBLICACIÓN

O paquete contén dúas funcións que permiten, de xeito doado, facer accesible o libro.

- *RStudio Connet:* creado para albergar o libro en formato HTML. Repositorio en https://bookdown.org.
- *GitHub:* Subir o libro ao GitHub

Referencias

[1] Yihui Xie (2018). bookdown: Authoring Books and Technical Documents with R Markdown. R package version 0.7.

[2] Allaire, J., Xie, \textit{*et al.*} (2018). Rmarkdown: Dynamic Documents for R. http://rmarkdown.rstudio.com, https://github.com/rstudio/rmarkdown.

[3] Yihui Xie (2016) *bookdown: Authoring Books and Technical Documents with R Markdown*. Champan and Hall/CRC.

[4] Yihui Xie, J. J. Allaire, Garrett Grolemund (2018). *R Markdown: The Definitive Guide.* Champan and Hall/CRC.

ESTIMACIÓN DE CONXUNTOS CON R MEDIANTE OS PAQUETES alphahull E alphashape3d

Beatriz Pateiro López[1]

[1] Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela

RESUMO

A estimación de conxuntos é unha rama da estatística que se enmarca dentro do campo da estatística non paramétrica, e na que ten especial importancia o emprego de ferramentas xeométricas. Ademais, ten aplicacións en campos moi diversos. Este traballo pretende facer unha revisión da implementación de ferramentas de estimación de conxuntos dispoñibles en R a través dos paquetes `alphahull` e `alphashape3d`.

Palabras e frases chave: estimación de conxuntos, alphahull, alphashape3d.

1. INTRODUCIÓN

En termos moi xerais, poderímos dicir que o obxectivo da estimación de conxuntos é aproximar a forma dun conxunto descoñecido a partir de observacións relacionadas con él. A estimación de conxuntos é unha rama da estatística, que se enmarca dentro do campo da estatística non paramétrica, e na que ten especial importancia o emprego de ferramentas xeométricas. Ademais, ten aplicacións en campos moi diversos como a bioloxía, a análise de imaxes, ou a econometría.
O paquete `alphahull` (ver [3]) calcula o α-shape e a envoltura α-convexa dunha mostra de puntos no plano. O concepto de α-shape e envoltura α-convexa xeneraliza a definición da envoltura convexa dun conxunto finito de puntos. A programación baséase na dualidade que existe entre o diagrama de Voronoi e a triangulación de Delaunay. O paquete inclúe tamén unha función que devolve ambas estruturas. O paquete `alphshape3d` (ver [4]) implementa en R o α-shape dun conxunto finito de puntos no espazo tridimensional. Desde a súa publicación, ambos paquetes contan cun número significativo de descargas e son citados en bibliografía de distintos campos de coñecemento.

2. APLICACIÓNS

Son moitas as áreas de aplicación da estimación de conxuntos. Por amosar soamente unha delas, centrarémonos neste apartado na estimación do "home range". A área de actividade dun animal (home range) é a área atravesada por un individuo durante as súas actividades normais de recolección de alimento, apareamiento, e coidado de crías [1]. Desde esta primeira definición, o concepto de área de actividade evolucionou, dando lugar a unha cantidade considerable de literatura sobre o tema, véxase por exemplo a revisión [5]. A estimación da área de actividade dun animal adóitase realizar a partir dun conxunto de localizacións do mesmo recollidas durante un período de tempo.
Son varias as metodoloxías propostas na literatura para a estimación da área de actividade, varias delas baseadas en técnicas de estimación de conxuntos.

Ademais, moitos destes métodos de estimación do home range xeralmente tratan as localizacións rexistradas como observacións independentes. Con todo, os avances recentes en tecnoloxía de seguimento animal (transmisión de radio do VHF, sistema Argos, e especialmente GPS) permítennos dispoñer dos movementos do animal prácticamente en tempo continuo. Neste contexto, a hipótese de independencia das observacións non ten sentido e precísanse novos modelos matemáticos. Así, [2] modeliza o movemento dun animal como un proceso estocástico continuo. A implementación desta situación tamén está contemplada no paquete alphahull (ver Figura 1).

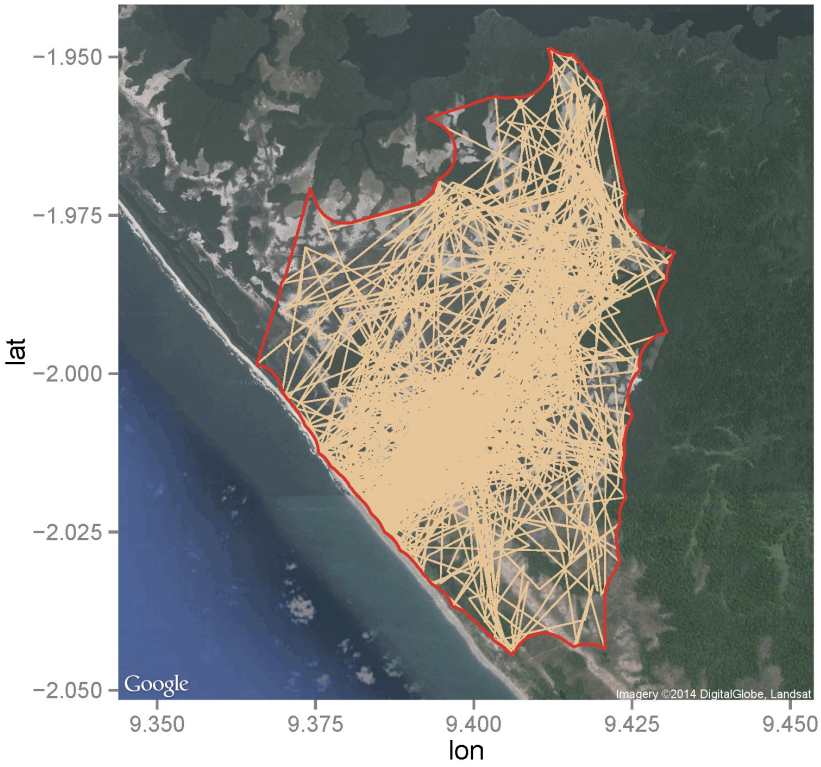


Figura 1: Estimación do home range do movimento dun elefante con n = 1633 posicións. En vermello, fronteira do estimador a-convex hull. Gráficas obtidas mediante o paquete alphahull.

AGRADECEMENTOS

Proxecto (MTM2016-76969P) finaciado polo Ministerio de Economía y Competitividad e cofinanciado por European Regional Development Fund (ERDF).

Referencias

[1] Burt, W. H. (1943) Territoriality and Home Range Concepts as Applied to Mammals. *J. Mammal.*, 24, 346-352.

[2] Cholaquidis, A., Fraiman, R., Lugosi, G. e Pateiro-Lopez, B. (2016) Set estimation from reected Brownian motion. *Journal of the Royal Statistical Society: Series B*, 78, 1057-1078.

[3] Pateiro-Lopez, B. e Rodrguez-Casal, A. (2010) Generalizing the Convex Hull of a Sample: The R Package alphahull. *J. Stat. Softw.*, 34(5), 1-28.

[4] Lafarge, T., Pateiro-Lopez, B., Possolo, A., Dunkers, Joy P (2014) R implementation of a polyhedral approximation to a 3D set of points using the -shape. *J. Stat. Softw.*, 56(4), 1-19.

[5] Powell, R. A. (2000) Animal home ranges and territories and home range estimators. In *Research Techniques in Animal Ecology: Controversies and Consequences* (eds Boitani, L. & Fuller, T.), pp 65-110. Columbia University Press, New York

MODELOS ESTADÍSTICOS DE CLASIFICACIÓN CON ALTA DIMENSIÓN EN EL NÚMERO DE COVARIABLES

Laura Freijeiro González[1]

[1] Universidade de Santiago de Compostela

RESUMEN

En esta presentación se hará una exposición de los problemas que aparecen a la hora de implementar las reglas de clasificación usuales del Análisis Discriminante en un contexto donde se cuenta con un mayor número de variables que de muestras. Se propondrán alternativas para afrontar este problema apoyándose en el programa de software libre R (R Core Team (2018)) y se mostrará el uso del mismo sobre una base de datos biomédicos (véase Guyon (2003)).

Palabras y frases clave: Alta dimensión, Análisis Discriminante.

1. EL ANÁLISIS DISCRIMINANTE

Las reglas de clasificación proporcionan un algoritmo que permite determinar en qué grupo clasificar una nueva observación dentro de las posibles $L \geq 2$ clases consideradas de antemano, teniendo únicamente en cuenta sus características o el valor de sus parámetros. De esta forma se tiene un procedimiento que permite extraer conclusiones en base a la discriminación realizada, como podría ser conocer si un paciente está sano o enfermo en base a su condición fisiológica o establecer en qué grupo de riesgo se encuentra un futuro cliente que quiere contratar un seguro. Para lograr este fin, el análisis discriminante se encarga de buscar las coincidencias y discrepancias entre dichas clases con el objetivo de dictaminar criterios que permitan entender qué hace único cada grupo y qué tipos de datos se asocian con estos, atendiendo a cómo será la estructura y forma de las fronteras de decisión.

Ante un escenario de alta dimensión donde se tiene un número de covariables mayor al tamaño muestral, $p>n$, las reglas estimadas de clasificación no poseen un buen comportamiento. Esto es debido a que tienen que afrontar problemas como el mal condicionamiento o el desastre de la dimensionalidad, distorsionando los resultados obtenidos y haciendo dudosa su eficiencia. En consecuencia será necesario recurrir a regularizaciones o modificaciones de estas para poder llevar a cabo la clasificación.

2. ANÁLISIS LINEAL DISCRIMINANTE (LDA) Y ANÁLISIS DISCRIMINANTE CUADRÁTICO (QDA)

Tanto la regla LDA (*Linear Discriminant Analysis*) o regla lineal de Fisher, como la QDA (*Quadratic Discriminant Analysis*) o regla discriminante cuadrática, se basan en que la variable poblacional sigue una distribución normal o Gaussiana y son las más eficientes en términos de conseguir la mejor separación entre clases, a través de fronteras lineales o cuadráticas respectivamente, Figura 1.

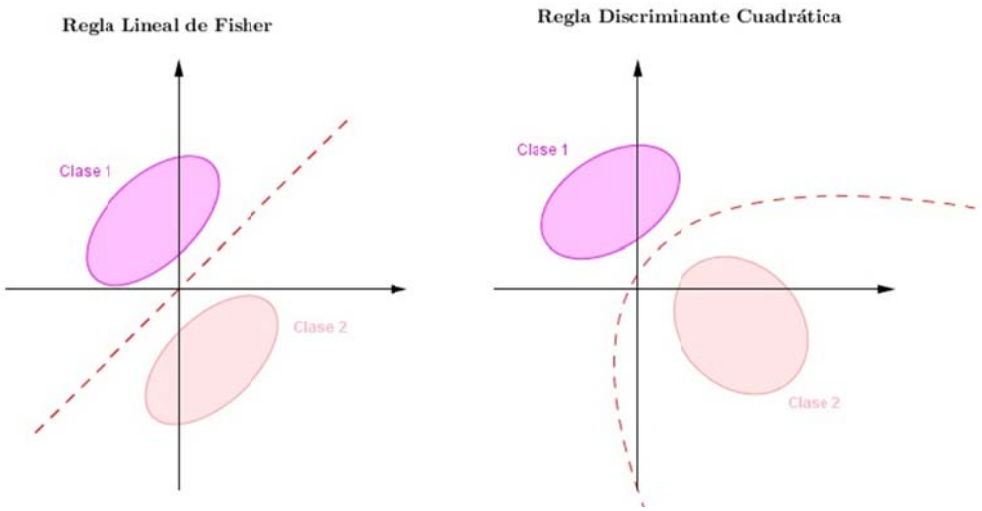


Figura 1: Gráfica de las fronteras de clasificación de las distintas reglas discriminantes en dos dimensiones.

El problema aparece a la hora de estimar estos criterios en el contexto de *p>n*, dado que las matrices de covarianzas Σ_j, con $j = 1, \ldots, L$ denotando las clases, son singulares y por lo tanto dichas reglas no pueden ser estimadas.

- **Regla LDA** supone que $\Sigma_j = \Sigma$ y clasifica en G_1 cuando:

$$\lambda^t \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right] > \log\left(\frac{\pi_2}{\pi_1} \right) \quad \text{siendo } \lambda = \Sigma^{-1}(\mu_1 - \mu_2)$$

- La **Regla QDA** asume que las matrices de covarianzas entre clases son distintas y clasifica un nuevo dato en el grupo que cumple

$$\max_l \delta_l(x) = -\frac{1}{2}\log(|\Sigma_l|) - \frac{1}{2}(x - \mu_l)^t \Sigma_l^{-1}(x - \mu_l) + \log(\pi_l)$$

Para solucionar este problema se puede recurrir a la utilización de versiones regularizadas de las matrices de covarianzas.

Se mostrará cómo se pueden llevar a cabo las versiones regularizadas de estos algoritmos empleando R, analizando las ventajas e inconvenientes que presenta dicho programa en este contexto.

3. REGRESIÓN LOGÍSTICA REGULARIZADA

Manteniendo el supuesto de normalidad en los datos, otra opción para solucionar este problema es recurrir a la **regresión logística regularizada**. Esta se basa en la construcción de un modelo regresión donde la variable respuesta *Y* es una **variable binaria (o dicotómica)**, es decir, sólo puede tomar dos valores. Se representarán estos valores por *0* y *1* respectivamente.

De esta forma se construye un modelo para la probabilidad de éxito condicionada,

$\pi(x) = \mathbb{P}(Y = 1 \mid X = x)$. Empleando la **función logit** $g(p) = \log\left(\frac{p}{1-p}\right) \quad \forall p \in [0,1]$, se obtiene:

$$\log\left(\frac{\pi(x,\beta)}{1-\pi(x,\beta)}\right) = x^t\beta.$$

De nuevo, el fenómeno de la alta dimensión forzará a que sea necesario trabajar con versiones regularizadas del modelo. En esta parte se mostrará la funcionalidad de la librería `glmnet` de Simon *et al.* (2011) para poder utilizar esta regla de clasificación, enseñando las distintas posibilidades que ofrece.

3. *K*-VECINOS MÁS CERCANOS Y *K*-MEDIAS

Otra forma de implementar discriminación es recurrir a método como el algoritmo de *K*-vecinos más cercanos o la regla de *K*-medias, Figura 2, los cuales no necesitan suponer ningún tipo de hipótesis distribucional.

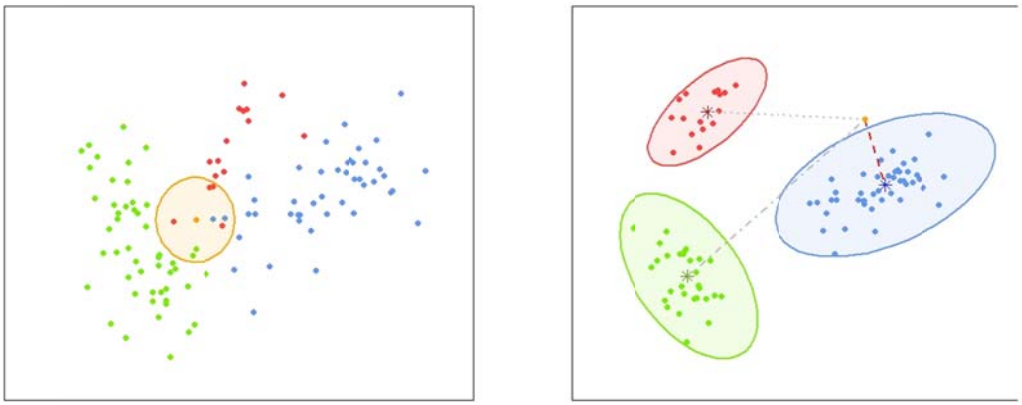


Figura 2: Ejemplo del método de k-vecinos más cercanos tomando k=8 en dos dimensiones (izquierda) y ejemplo del algoritmo de k-medias para el caso k=3 (derecha).

Estos métodos son fáciles de emplear en R, en particular se han escogido las librerías `class` (Venables y Ripley (2002)) para el algoritmo de *k*-vecinos y la librería `stats` (R Core Team (2018)) para el método de *k*-medias por su sencillez y eficacia. En este apartado se mostrará el funcionamiento de las mismas.

4. SUPPORT VECTOR MACHINE (SVM)

Finalmente se expondrá la utilización de los SVM (*Support Vector Machines*) para solucionar el problema de clasificación cuando *p>n*. Este método se basa en resolver un problema de optimización con la finalidad de conseguir construir una banda lo más amplia posible para separar las clases, contemplando tanto los casos donde los grupos estén completamente separados como donde pueda existir solapamiento, Figura 3.

- **Clases sin solapamiento**

$$\begin{cases} \min_{\beta,\beta_0} \|\beta\| \\ \text{sujeto a } y_i(x_i^t\beta+\beta_0)\geq 1, \quad i=1,\ldots,n. \end{cases}$$

- **Clases con solapamiento**

$$\begin{cases} \min \|\|\beta\|\| \text{ sujeto a} & y_i(x_i^t\beta+\beta_0)\geq 1-\xi_i \quad \forall i, \\ & \xi_i\geq 0, \quad \sum\xi_i\leq cte \end{cases}$$

- **Regla discriminante:**

$$\hat{G}(x)=\text{sign}[\hat{f}(x)]=\text{sign}[x^t\hat{\beta}+\hat{\beta}_0]$$

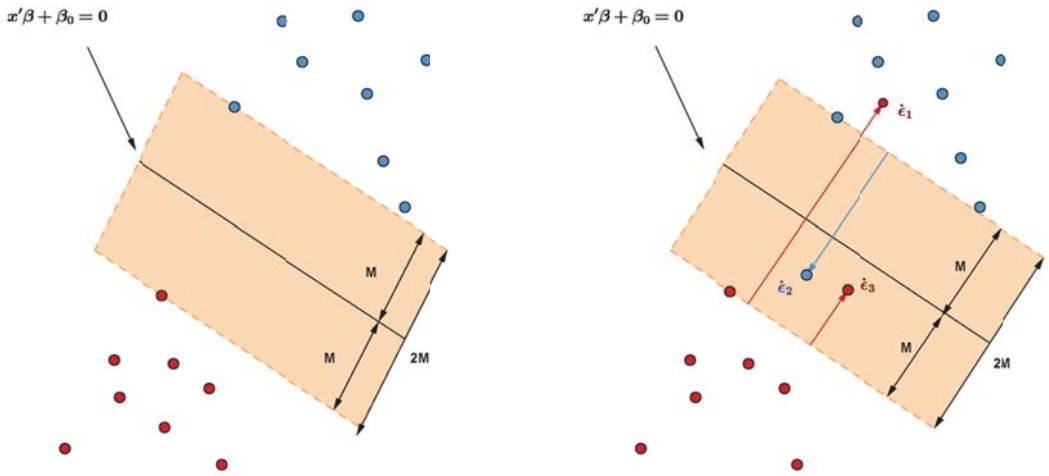


Figura 3: Ejemplo del clasificador SVM en dos dimensiones con clases sin solapamiento (izquierdo) y considerando solapamiento (derecha).

Se mostrará como emplear este método a través de la librería `e1071` (Meyer et al. (2018)) para ambos contextos, viendo las opciones que proporciona y los resultados obtenidos.

Referencias

[1] Guyon, I. (2003). *Design of experiments for the NIPS 2003 variable selection benchmark.*

[2] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., y Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group* (Formerly: E1071), TU Wien. R package version 1.7-0.

[3]R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[4] Simon, N., Friedman, J., Hastie, T., y Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1-13.

[5]Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

COMPARANDO MÉTODOS DIAGNÓSTICOS EN R

Arís Fanjul Hevia[1], Wenceslao González Manteiga[1] y Juan Carlos Pardo Fernández[3]

[1] Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

[2] Departamento de Estatística e Investigación Operativa, Universidade de Vigo

RESUMEN

Una de las cuestiones más importantes en cualquier estudio médico es el ser capaz de diagnosticar correctamente a los pacientes afectados por una determinada enfermedad. Un método de diagnosis es, en realidad, un problema de clasificación en el que se trata de minimizar el número de enfermos a los que no se les detecta la enfermedad, y el número de sanos que son declarados enfermos.

Una forma usual de analizar el buen comportamiento de un método de diagnosis es estudiar su curva ROC (del inglés, Receiver Operating Characteristic). Esta herramienta estadística utiliza las nociones de sensibilidad y especificidad para analizar la capacidad discriminativa un método de clasificación [2].

En particular, comparar curvas ROC correspondientes a distintos métodos de diagnosis puede servir para determinar si esos métodos son equivalentes o si, por el contrario, uno de ellos es capaz de clasificar de forma más precisa a los individuos en estudio. Existen varias librerías en R capaces de estimar, dibujar y comparar estas curvas ROC, como puede ser `pROC` [4] o `nsROC` [3].

Por otra parte, es usual que en el proceso de diagnosis se cuente con información extra de covariables, información que es importante incluir en el estudio porque puede influir en la capacidad discriminativa de las curvas ROC. Esto se puede hacer considerando la curva ROC condicionada [1].

En este trabajo se estudia una base de datos real de enfermos con derrame pleural sospechosos de tener cáncer. Se verá qué herramientas existen ya en R para analizar la capacidad discriminativa de una variable diagnóstica, y se mostrarán otras nuevas para estimar una curva ROC condicionada a una covariable. Finalmente se aplicará una nueva metodología para comparar curvas ROC condicionadas, y se verá que, si no se tiene en cuenta esta información extra, se puede llegar a una conclusión diferente.

Palabras y frases clave: Bootstrap, comparación, covariables, curvas ROC, estimación.

AGRADECIMIENTOS

Los autores agradecen el soporte económico del Ministerio de Educación, Cultura y Deporte a través de la ayuda FPU14/05316 y el soporte económico del Ministero de Economía, Industria y Competitividad a través de los proyectos MTM2016-76969-P, MTM2014-55966-P y MTM2017-89422-P, que incluyen apoyos del European Regional

Development Fund y la Agencia Estatal de Investigación. También se agradece al Prof. F. Gude (Unidad de Epidemiología Clínica del Hospital Clínico Universitario de Santiago) por proporcionar los datos del estudio.

Referencias

[1] González-Manteiga, W., Pardo-Fernández, J.-C., and Van-Keilegom, I. 2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*}, 38(1):169-184.

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

[3] Pérez-Fernández S (2018) nsROC: Non-Standard ROC Curve Analysis. R package version 1.1. https://CRAN.R-project.org/package=nsROC.

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

FERRAMENTAS PARA REDUCIR O TEMPO DE EXECUCIÓN EN R

Alejandra López Pérez[1]

[1] Universidade de Santiago de Compostela

RESUMO

R (R Core Team, 2018) é unha linguaxe e entorno de programación para a análise estatística e gráfica. É unha linguaxe interpretada (ou de *implementación interpretada*), polo que non se caracteriza pola súa velocidade. Non obstante, existen ferramentas para incrementar a velocidade de cálculo dos nosos scripts de R, que introduciremos neste relatorio. Entre outras solucións cubrirase a vectorización de código, a compilación a bytecode, a interfaz de programación de aplicacións (API) para extender R con código escrito en C ou Fortran, o paquete `Rcpp` (Eddelbuettel e Balamuta, 2017) para a integración de R e C++, `gpuR` (Rupp et al., 2016) para a computación na GPU, e `data.table` (Dowle e Srinivasan, 2018), unha extensión do obxecto `data.frame`. Además, para aquelas tareas susceptibles de paralelización, R dispón de diversos paquetes, como `parallel` e `doParallel` (Corporation e Weston, 2017), que permiten empregar varios procesadores ao mesmo tempo.

Palabras e frases chave: parallel computing, bytecode, compiled language, Rcpp.

AGRADECEMENTOS

Debo agradecer ao Ministerio de Economía, Industria y Competitividad a concesión da beca FPI MTM2016-76969-P.

Referencias

[1] Corporation, M. eWeston, S. (2017). *doParallel: Foreach Parallel Adaptor for the 'parallel'Package*. R package version 1.0.11.

[2] Dowle, M. e Srinivasan, A. (2018). *data.table: Extension of 'data.frame'*. R package version 1.11.0.

[3] Eddelbuettel, D. e Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, 5:e3188v1.

[4] R Core Team (2018). R: A language and environment for statistical computing.

[5]Rupp, K., Tillet, P., Rudolf, F., Weinbub, J., Grasser, T., e J ungel, A. (2016). Viennacl-linear algebra library for multi- and many-core architectures. *SIAM Journal on Scientific Computing*.

UNHA APLICACIÓN SHINY DE R PARA A XESTIÓN DE RECURSOS EN INCENDIOS FORESTAIS

Jorge Rodríguez-Veiga[1], María José Ginzo-Villamayor[2] e Balbina Casas-Méndez[2]

[1] Instituto Tecnológico de Matemática Industrial (ITMATI)
[2] Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela (USC)

RESUMO

Determinar a planificación óptima que permita decidir o tipo e o número de recursos necesarios para extinguir un incendio forestal é unha tarefa que se abordou na literatura utilizando modelos extraídos da investigación operativa. Neste traballo proponse un modelo de optimización que contempla tamén a asignación dos recursos a diferentes períodos de tempo. Esta asignación resulta de interese para os profesionais que combaten os incendios, que deben de cumprir coa normativa española que regula os períodos de traballo e descanso dos pilotos e brigadas, sen descoidar a restrición natural de non desatender os frontes máis perigosas do incendio.

Por último creamos unha interface gráfica que facilite a introdución dos datos e proporcione os resultados de maneira práctica. A programación informática, do modelo e da interface, realizámola coa linguaxe R.

Palabras e frases chave: supresión de incendios forestais; planificación de recursos; asignación de tempo; programación lineal enteira; interface; shiny.

1. INTRODUCIÓN

O fenómeno dos incendios forestais converteuse nun dos principais problemas ecolóxicos que sofren os bosques, debido á alta frecuencia e intensidade que adquiriron nas últimas décadas e que leva graves consecuencias económicas (Butry *et al.*, 2001). Cada ano prodúcense máis de 45.000 incendios no sur de Europa, que teñen como consecuencia a queima de preto de 0,5 millóns de hectáreas de bosques e outras terras rurais (Camia *et al.*, 2009). Os períodos de seca ou o cambio climático non fan máis que acrecentar este fenómeno (Arno e Allison- Bunnell, 2002), e aínda que o esforzo para previr incendios é importante (Martínez *et al.*, 2009), é esencial contar con ferramentas que permitan unha toma de decisións eficiente cando o lume xa se produciu e debe ser contido (Mavsar *et al.*, 2013, Martell, 2015). En Galicia (cunha superficie de 29.574 km^2, o 69% da mesma está formada por bosques), o organismo público rexional responsable de combater incendios forestais conta con 30 avións en 2017, dos cales 25 son helicópteros. É importante unha boa planificación dos recursos de extinción que reduza os custos e danos causados polo incendio. A análise dos incendios desde a perspectiva da optimización dos custos asociados remóntase a Headley (1916) e Sparhawk (1925). O modelo teórico que persegue unha xestión de recursos baseada na minimización de custos, en sentido amplo, foi denominado C+ NVC (*Cost Plus Net Value Change*, Gorte e Gorte, 1979). Este modelo combina o obxectivo de minimizar o custo provocado polo uso de recursos (terrestres e/ou aéreos) co de reducir os custos xerados pola queima do terreo, a perda de materiais

ou as tarefas de rexeneración. Donovan e Rideout (2003a) propoñen unha reformulación do modelo inicial e Donovan e Rideout (2003b) propoñen un modelo de programación lineal determinista que minimiza a función C+ NVC e permite seleccionar os recursos para utilizar considerando o momento en que se pode conter o incendio cos recursos dispoñibles.

2. MARCO TEÓRICO E ENFOQUE METODOLÓXICO

Propoñemos un modelo de programación lineal binario cuxa meta é seleccionar os recursos aéreos que se utilizarán, durante un día de traballo, para a extinción dun incendio forestal. Ao mesmo tempo, proporciona unha planificación temporal dos recursos que se axusta aos tempos máximos de voo e os tempos normativos de descanso. A función obxectivo do modelo, seguindo a metodoloxía C+ NVC, contempla minimizar o tempo de extinción de incendios e, en consecuencia, minimizar os danos causados polo lume e o tempo de utilización dos recursos.

O modelo foi programado co software libre R (The R Project for Statistical Computing, 2017) e a súa resolución foi contrastada mediante o solver Gurobi (Gurobi Optimizer, 2017) por medio de distintos datos históricos.

Finalmente, neste traballo móstrase unha interface construída coa ferramenta R que permite utilizar o algoritmo dunha maneira sinxela e xerar informes das actividades relacionadas coa planificación dos recursos.

3. FORMULACIÓN DO PROBLEMA

A formulación matemática realizada modela a función obxectivo e as limitacións que se impoñen para identificar a asignación óptima no problema da selección de recursos para a contención dun incendio forestal.

O modelo contempla unha serie de conxuntos: conxunto das posibles aeronaves, brigadas a seleccionar para combater o lume e o conxunto de períodos de tempo nos que se ten a información da evolución do lume. Tamén unha serie de parámetros: neste caso distinguiremos entre aqueles relativos as aeronaves, as brigadas, ao lume e os relativos á normativa.

A función obxectivo plantexada minimiza a suma dos custos involucrados na extinción do incendio forestal. Ten en conta os seguintes sumandos:

1. o custo variable por utilizar os recursos seleccionados,
2. o custo fixo asociado á utilización de cada un deles e
3. o producido polas hectáreas de terreo queimado, que se inclúe para penalizar o incumprimento no número mínimo de aeronaves.

O modelo está suxeito a unha serie de restricións:
As relativas a contención do lume; relativas ao inicio e fin da actividade; sobre os períodos de descanso; sobre o tempo de voo das aeronaves; número de aeronaves e número de brigadas.

Todos os detalles do modelo poden consultarse en Rodríguez-Veiga *it et al*., 2018.

4. INTERFACE

Co fin de facilitar a execución do modelo creouse unha interface co software R mediante o uso da librería `shinydashboard`.

A interface permite unha sinxela interacción á hora de cargar a información das aeronaves así como da evolución do incendio e da normativa española de aviación. Ademais ofrece a opción de executar os modelos con distintos solvers (*gurobi*, *lpSolve* e *Rsymphony*, estes dous últimos de licenza libre). Unha vez realizada a execución, móstrase unha visualización dos resultados tanto gráfica (mediante gráficos interactivos) como en forma de táboas e permítese a creación de *informes* para gardar a información da instancia executada nun arquivo en formato html.

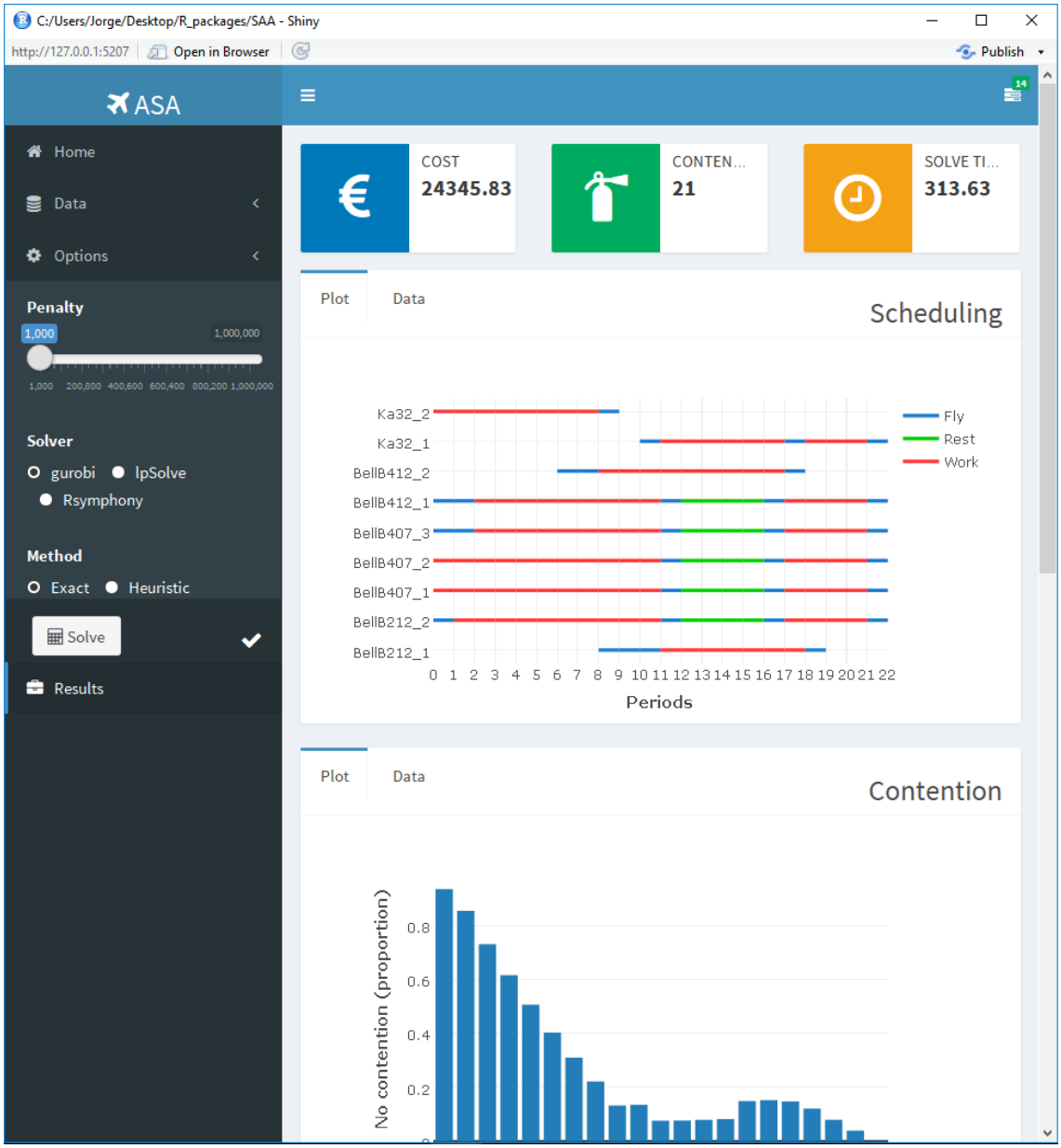


Figura 1: Resolución do modelo de asignación de recursos coa interface creada con Shiny.

Na Figura 1 móstrase o resultado de resolver unha instancia na que se dispón de 9 aeronaves e coñécese a evolución do incendio durante 25 períodos de tempo, sendo cada un deles de 10 minutos, co que temos a evolución do incendio nos 250 minutos posteriores. Tense en conta ademais, a normativa regulada pola Circular Operacional 16--B (1995). Como vemos, obtense que o incendio se conterá no período 21 (aos 210 minutos) cun custo de operación de 24.345,83€. A xanela *scheduling* móstranos para

cada aeronave seleccionada a súa planificación horaria que inclúe os tempos de traballo, de voo e de descanso. Observamos, en particular, que en ningún período de tempo o incendio queda desatendido. O tempo que durou a execución foi de 314 segundos.

AGRADECEMENTOS

Os investigadores agradecen o apoio financeiro recibido desde o Ministerio de Economía y Competitividad de España a través dos proxectos MTM2014-53395-C3-2-P e MTM2016-76969-P, e desde ITMATI, a través do proyecto Enjambre.

Referencias

[1] Arno, S. F. and Allison-Bunnell, S. (2002) Flames in our forest: disaster or renewal? *Island Press*, Washington, DC.

[2] Butry, D. T., Mercer, D. E., Prestemon, J. P., Pye, J. M., and Holmes, T. P. (2001) What is the price of catastrophic wildfire? *Journal of Forestry* 99, 9-17.

[3] Camia, A., San-Miguel-Ayanz, J., Oehler, F., Santos De Oliveira, S., Durrant Houston, T., Kucera, J., Boca, R., Whitmore, C., Giovando, C., Amatulli, G., Libertà, G., Schmuck, G., Schulte, E., and Bucki, M. (2009) Forest Fires in Europe 2008. EUR 23971 EN. Luxembourg (Luxembourg): OPOCE; 2009. JRC53463.

[4] Donovan, G. and Rideout, D. (2003a) A reformulation of the Cost Plus Net Value Change (C + NVC) model of wildfire. *Forest Science* 49, 318-323.

[5] Donovan, G. and Rideout, D. (2003b) An integer programming model to optimize resource allocation for wildfire containment. *Forest Science* 49, 331-335.

[6] Gorte, J. K. and Gorte, R. W. (1979) Application of Economic Techniques to Fire Management-A status Review and Evaluation. *USDA Forest Service General Technical Report* INT-53.

[7] Gurobi Optimizer (2017). Available online at: http://www.gurobi.com/. Last October 1, 2018.

[8] Headley, R. (1916) Fire Suppression, District 5. United States Departament of Agriculture, Forest Service, Washington, 57 p.

[9] Martell, D. L. (2015) A review of recent forest and wildland fire management decision support systems research. *Current Forestry Reports* 1, 128-137.

[10] Martínez, J., Vega-García, G., and Chuvieco, E. (2009) Human-caused wildfire risk rating for prevention planning in Spain. *Journal of Environmental Management* 90, 1241-1252.

[11] Mavsar, R., González-Cabán, A., and Varela, E. (2013) The state of development of fire management decision support systems in America and Europe. F*orest Policy and Economics* 29, 45-55.

[12] Rodríguez Veiga, J., Ginzo-Villamayor, M.J., Casas-Méndez, B. V. (2018). An Integer Linear Programming Model to Select and Temporally Allocate Resources for Fighting Forest Fires *Forests* (section: Forest Ecology and Management). 9, 583. pp. 1-18; doi:10.3390/f901000583. MDPI.

[13] Sparhawk, W. N. (1925) The use of liability ratings in planning forest fire protection. *Journal of Agricultural Research* 30, 693-792.

[1] The R Project for Statistical Computing (2017). Available online at:

https://www.r-project.org/. Last accessed October 1, 2018.

OBRADOIRO: INTRODUCIÓN AO SOFTWARE R

Mª José Ginzo Villamayor[1]

[1] Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela (USC)

RESUMO

Neste obradoiro mostrarase unha introducción a linguaxe de programación R (importar datos, obxectos en R, operacións básicas, vectores, matrices, data frame, as funcións outer e apply entre outras, ...), así como os diferentes tipos de representacións gráficas que se poden facer, principáis liñas de código ou comandos para facer unha sinxela análise estatística e presentación do paquete R Commander (rcmdr).

Requirimentos, ter descargado no PC:

- R e
- Rstudio.

OBRADOIRO: INICIACIÓN Ó BIG DATA CON R COA LIBRARÍA sparklyr

M. Aurora Baluja González[1], Javier López Cacheiro[2]

[1] Servizo de Anestesioloxía e Reanimación. Complexo Hospitalario Universitario de Santiago (CHUS)

[2] Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia (CESGA)

RESUMO

A tecnoloxía que permite a lectura e escritura paralela e eficiente de macrodatos ("Big Data") experimentou unha gran difusión nos últimos anos, coa aparición de plataformas open source - Hadoop, Spark-, e linguaxes como Scala.

O paquete sparklyr, lanzado en 2016 por RStudio, permite unha comunicación directa coa API de Spark para " dataframes2", utilizando a sintaxe de R e dplyr, de forma cómoda para o usuario final.

Neste obradoiro aprenderemos:

- Os conceptos máis importantes sobre o funcionamiento da tecnoloxía para Big Data.
- A posta en marcha dunha instancia de Spark no noso PC, ou (para os que teñan conta) no servidor Big Data do CESGA.
- As operacións máis frecuentes que permite o stack Hadoop-Spark-Sparklyr sobre os nosos datos.

Requírense coñecementos básicos de R.
Para o obradoiro requírese ter instalado e en funcionamento (apórtanse suxerencias de vencellos de axuda) :

1. Pasos comúns aos 3 sistemas operativos:
   - Instalar R: https://cran.r-project.org/.
   - Instalar R Studio: https://www.rstudio.com/products/rstudio/download/#download
   - Instalar algunha ferramenta que permita usar Git (recomendable, non imprescindible).
2. MS Windows:
   - Instalar Java (JDK): https://www.theserverside.com/tutorial/How-to-install-the-JDK-on-Windows-and-setup-JAVA_HOME.
   - Instalar Apache/Spark: https://hernandezpaul.wordpress.com/2016/01/24/apache-spark-installation-on-windows-10/.
   - Instalar paquete sparklyr desde R.

3. OSX (o proceso pode levar bastante tiempo no caso de se completar desde cero)

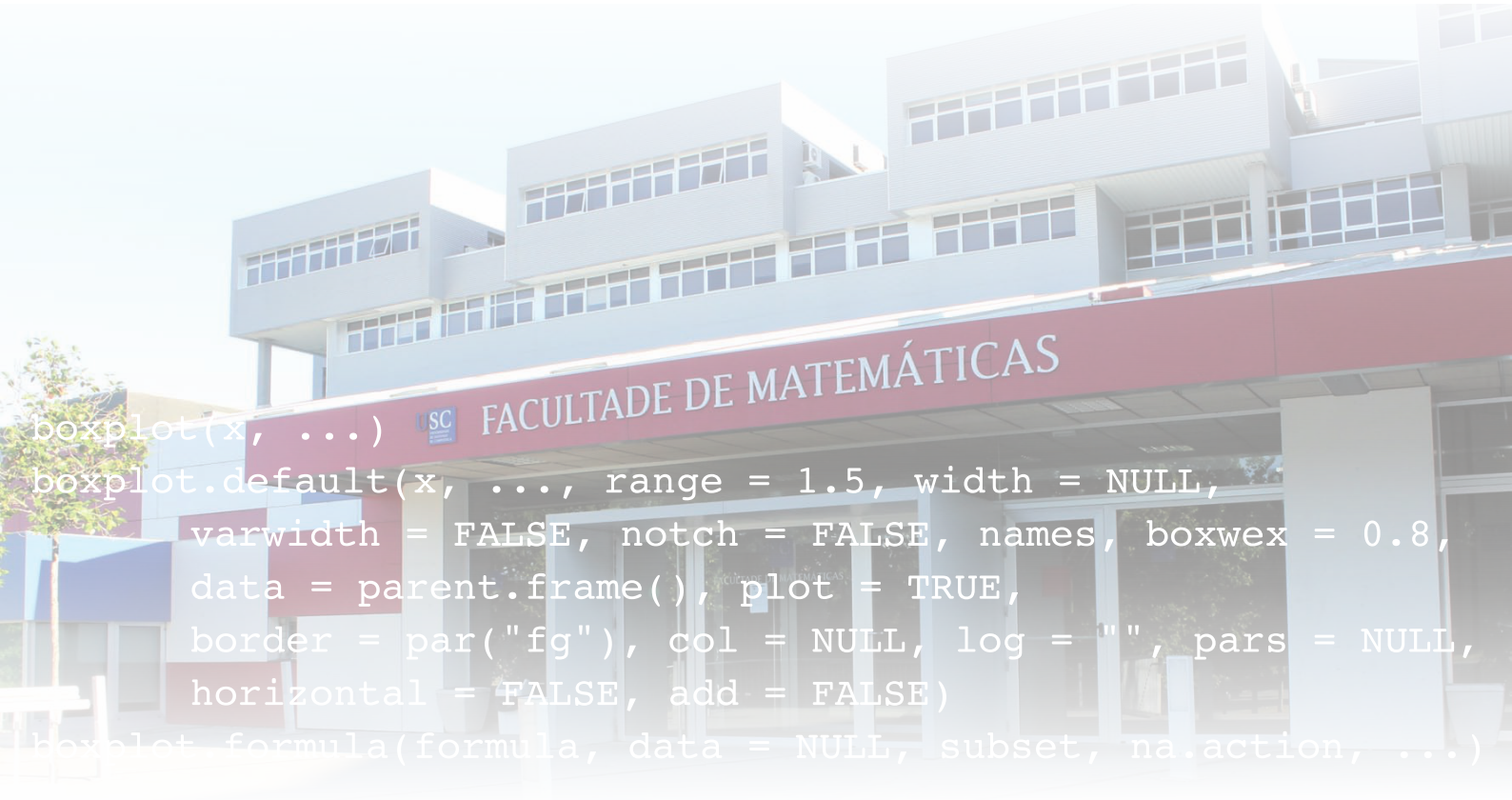
- Instalar xcode: https://developer.apple.com/xcode/ (a descarga e instalación poden levar tempo, ó instala-las command-line developer tools)
- Instalar Java e Apache/Spark: https://medium.freecodecamp.org/installing-scala-and-apache-spark-on-mac-os-837ae57d283f.
- Instalar paquete sparklyr desde R.

4. GNU/Linux (exemplo con Ubuntu 16.04):

- Instalar Java e Apache/Spark: https://www.tutorialkart.com/apache-spark/install-latest-apache-spark-on-ubuntu-16/.
- Instalar paquete sparklyr desde R.

AUTORES

V XORNADA DE USUARIOS DE R EN GALICIA



> ORGANIZA

CIXUG

> COLABORAN

R-Galicia

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE MATEMÁTICAS

> PATROCINAN

amtega Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

ISBN 9 788409 058051 >