VII XORNADA DE USUARIOS DE R EN GALICIA

15 de outubro de 2020

LIBRO DE RESUMOS

> ORGANIZA

CIXUG

> COLABORA

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

VII XORNADA DE USUARIOS DE R EN GALICIA

PROGRAMA E RESUMOS

15 de outubro de 2020

Organiza: Oficina de Software Libre do CIXUG

VII XORNADA DE USUARIOS DE R EN GALICIA

Editora: María José Ginzo Villamayor

ISBN: 978-84-09-24273-3

Presentación

A Oficina de Software Libre (OSL) do CIXUG comprácese en presentar a VII 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 dezaoito relatorios ao longo de todo o día. Dos cales catro son convidados e ás outras catorce atenderon á chamada de recepción de propostas. Durante a xornada teremos un espazo adicado a traballos relacionados co COVID-19, tan protagonista neste 2020, que aínda así, grazas ás novas tecnoloxías nos permite levar a cabo esta xornada.

Entre as persoas participantes figuran especialistas do Instituto Español de Oceanografía (IEO), da Consellaría de Sanidade, das tres universidades galegas, así como doutras nacionais como son á Universidad de Castilla-La Mancha (UCLM), Universidad Miguel Hernández (UMH), ou internacionais como Universidad Cooperativa de Colombia, Universidad Santo Tomás, Universidad de Los Llanos, e dos seguintes centros de investigación: Centro de Investigación de Tecnoloxías da Información e das Comunicacións (CITIC), Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Centro Tecnolóxico de Telecomunicaciones de Galicia (Gradiant) e o Centro de Investigación Forestal (CIFOR) do INIA, o Instituto Tecnolóxico de Matemática Industrial (ITMATI) e de empresas como TasteLab & SENSESBIT e alumnos do programa de Doutoramento en Estatística e Investigación Operativa e do Máster en Técnicas Estatísticas.

Todo isto non sería posible sen o patrocinio da AMTEGA e a colaboración da Asociación de Usuarios de Software Libre da Terra de Melide (MeLiSA), ás que agradecemos a súa contribución.

Santiago de Compostela, outubro de 2020

O Comité Organizador

Comité organizador

María José Ginzo Villamayor
Universidade de Santiago de Compostela

Rafael Rodríguez Gayoso
Concello de Santiago de Compostela

Xabier Sánchez Santos
Consorcio Interuniversitario CIXUG

Comité científico

María José Ginzo Villamayor
Universidade de Santiago de Compostela

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

Información xeral

VII XORNADA DE USUARIOS DE R EN GALICIA

Emisión en vivo

Sala de Zoom

YouTube

Data

15 de outubro de 2020

Web das xornadas

https://www.r-users.gal/

Certificados

Todos os certificados remitiranse ás persoas solicitantes en formato dixital por correo electrónico unha vez rematada a VII Xornada.

Programa

15 de outubro de 2020

VII XORNADA DE USUARIOS DE R EN GALICIA

09:00	**Sesión de apertura** Xabier Sánchez Santos[1], María José Ginzo Villamayor[2], Rafael Rodríguez Gayoso[3] [1]Consorcio CIXUG e Comité Organizador, [2]Comité Científico da VII Xornada, [3]Asociación de Software Libre de Terras de Melide (MeLisa)
09:20	**Determinación de la calidad sensorial de alimentos mediante mapas de preferencia** Andrés Martínez Sánchez TasteLab & SENSESBIT
09:40	**Monitorización automática de incendios** Manuel Novo Pérez[1], Manuel Vaamonde Rivas[1], Marta Rodríguez Barreiro[1] e María José Ginzo Villamayor[2] [1]Instituto Tecnolóxico de Matemática Industrial (ITMATI), [2]Depto. de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela
10:00	**Cálculo de mapas de risco e ocorrencia de incendios** Marta Rodríguez Barreiro[1], Manuel Novo Pérez[1], Manuel Vaamonde Rivas[1] e María José Ginzo Villamayor[2] [1]Instituto Tecnolóxico de Matemática Industrial (ITMATI), [2]Depto. de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela
10:20	**RedBee: sistemas de análisis inteligente de ciberataques en honeypots** Marta Sestelo[1,2], Lilian Adkinson Orellana[1], Borja Pintos Castro[1], Cristian Marques Corrales[1], Iago López Román[1] e Nora M. Villanueva[1,2] [1]Gradiant, Centro Tecnológico de Telecomunicaciones de Galicia, [2]Depto. De Estatística e Investigación Operativa, Grupo SIDOR, Universidade de Vigo
10:40	**A new R-package for the analysis of the fisheries population under uncertainty** Marta Cousido-Rocha, S. Cerviño, M.G. Pennino Instituto Español de Oceanografía (IEO)
11:00	PAUSA
11:40	**Modelos de riscos competitivos – Modelos de riesgos competitivos** Aurora Baluja González Servizo de Anestesioloxía e Reanimación, Hospital Clínico Universitario de Santiago de Compostela
12:00	**Desarrollo de un atlas de mortalidad en Castilla-La Mancha con R** Virgilio Gómez Rubio[1] [1]Departamento de Matemáticas, Universidad de Castilla-La Mancha (UCLM)
12:20	**O emprego de R na detección das características máis influentes na clasificación de pacientes infectados por COVID-19 en Galicia** Laura Davila-Pena[1], Balbina Casas-Méndez[2] e Ignacio García Jurado[2] [1]IMAT, Instituto de Matemáticas, MODESTYA, Grupo de Modelos de Optimización, Decisión, Estatística e Aplicacións, Depto. de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, [2]CITIC, Centro de Investigación en Tecnoloxías da Información e das Comunicacións, MODES, Grupo de Modelización, Optimización e Inferencia Estatística, Depto. de Matemáticas, Facultade de Informática, Universidade da Coruña
12:40	**Functional regression models for the prediction of COVID-19** Manuel Oviedo de la Fuente[1] e Manuel Febrero Bande[2] [1]CiTIUS, Universidade de Santiago de Compostela, [2]Depto. de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela
13:00	**Emprego de R na pre dición cooperativa de variables relacionadas coa pandemia do COVID-19** Rubén Fernández Casal[1], Carlos Fernández Lozano[2] e José A. Vilar Fernández[1] [1]CITIC, Grupo MODES, Dept. de Matemáticas, Universidade da Coruña, [2]CITIC, Grupo RNASA-IMEDIR, Depto. de Ciencias da Computación e Tecnoloxías da Información, Universidade da Coruña
13:20	**FORTLS: un paquete de R para a estimación de variables dasométricas para o seu uso en inventario forestal** Juan Alberto Molina-Valero[1], María José Ginzo Villamayor[2], Manuel Antonio Novo Pérez[3], Juan Gabriel Álvarez-González[1], Fernando Montes[4] e César Pérez-Cruzado[1] [1]Unidade de Xestión Ámbiental e Forestal Sostible (UXAFORES), Dpto. de Enerxía Agroforestal, Escola Politécnica Superior de Enxeñería, Universidade de Santiago de Compostela, [2]Dpto. de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, [3]Instituto Tecnolóxico de Matemática Industrial (ITMATI), [4]INIA-CIFOR
13:40	**Visualización de la dirección fluvial** Dominic Royé Departamento de Xeografía, Universidade de Santiago de Compostela, Grupo de Epidemioloxía e Saúde Pública
14:00	PAUSA
16:00	**Nonparametric estimation of directional highest density regions** Paula Saavedra-Nieves[1] e Rosa María Crujeiras[1] [1]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela
16:20	**Estimación non paramétrica da densidade no plano con R** María Bugallo Porto Estudante do Mestrado en Técnicas Estatísticas, Universidade de Santiago de Compostela
16:40	**Análisis bibliométrico básico con Bibliometrix: el caso de la longitud del telómero en niños** Daniel Prieto-Botella[1], Desirée Varela -Gran[1,2], Paula Fernández-Pires[1], Paula Peral-Gómez[1,2], Miriam Hurtado-Pomares[1,2], Alicia Sánchez-Pérez[1,2], Iris Juárez-Leal[1,2], Cristina Espinosa-Sempre[1,2], Eva María Navarrete-Múñoz[1,2] [1]Depto. de Cirugía y Patología, Universidad Miguel Hernández, [2]Grupo de Investigación en terápia Ocupacional (InTeO), Universidad Miguel Hernández
17:00	**El R-kward en el análisis de modelos de regresión con datos del COVID-19 en Colombia** Jorge Alejandro Obando Bastidas[1], Laura Nathalia Obando[2] e María Teresa Costélanos Sánchez[3] [1]Universidad Cooperativa de Colombia, [2]Universidad Santo Tomás, [3]Universidad de Los Llanos
17:20	**R tools to link game theory and statistics by sampling** Alejandro Saavedra-Nieves Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela
17:40	**ProjectManagemet: un paquete de R para a xestión de proxectos** Juan Carlos Gonçalves-Dosantos[1], Ignacio García Jurado[1] e Julián Costa[2] [1]Grupos MODES, CITIC e Departamento de Matemáticas, Universidade da Coruña, [2]Grupo MODES, Departamento de Matemáticas, Universidade da Coruña

Índice

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

MODELOS DE RISCOS COMPETITIVOS - MODELOS DE RIESGOS COMPETITIVOS

Aurora Baluja[1]

[1]Servicio de Anestesioloxía e Reanimación. Hospital Clínico Universitario de Santiago de Compostela. España.

RESUMO

Los riesgos competitivos (RC) son eventos cuya ocurrencia excluye o altera la probabilidad de la ocurrencia de otro evento. Las situaciones de riesgos competitivos se dan frecuentemente cuando estudiamos un evento cualquiera, en una cohorte que puede estar sometida a un evento terminal (como por ejemplo el fallecimiento). Veremos cómo adecuar las variables de nuestro estudio para poder emplear modelos RC desde R.

Palabras e frases chave: competitivos, riscos, supervivencia, eventos.

1. INTRODUCIÓN

Un riesgo competitivo (RC) es un evento cuya ocurrencia excluye la ocurrencia de otro evento, o bien altera fundamentalmente la probabilidad de ocurrencia de este otro evento. Las situaciones de riesgos competitivos se dan frecuentemente cuando estudiamos un evento cualquiera en una cohorte que puede estar sometida a un evento terminal (como por ejemplo el fallecimiento).
Si no los tenemos en cuenta, el riesgo atribuído a un evento estará sesgado debido a que los individuos que experimentaron antes el riesgo competitivo tienen ahora un riesgo muy diferente (incluso nulo en el caso de muerte) de experimentar el evento de interés.

2. MODELOS DE COX DE RIESGOS COMPETITIVOS

El 'riesgo de causa específica', o 'cause-specific hazard'(CSH) en el tiempo t (modelo de riesgos competitivos de Cox) se define por el riesgo instantáneo del evento por unidad de tiempo debido a un riesgo, como si no hubiera eventos compitiendo. Sin embargo, los hazard ratio (HR) obtenidos no están directamente relacionados con la predicción de la incidencia acumulada del evento, por lo que no son adecuados para predecir el riesgo **global** de un paciente determinado.
La función de riesgo de causa específica es la siguiente:

$$h_k(t) = \lim_{\Delta \to 0} \frac{P(t < T < T + \Delta t, \delta = k | T > t}{\Delta t}, k = 1, ..., K$$

3. MODELOS FINE-GRAY DE RIESGOS COMPETITIVOS

La 'Hazard Ratio de la subdistribución'(modelo Fine-Gray) se interpreta como la probabilidad de observar un evento de interés en el siguiente intervalo de tiempo, sabiendo que el evento de interés no ocurrió hasta entonces o que se observó un evento competitivo. Desafortunadamente, los valores absolutos de los coeficientes de regresión en el modelo Fine-Gray no tienen una interpretación directa.
El cociente de riesgo de subdistribución denota la dirección, pero no proporciona directamente la magnitud del efecto de la covariable sobre la incidencia acumulada (CIF).

La función de incidencia acumulada (CIF), teniendo en cuenta un evento de interés (1) y su evento competitivo (2), se puede representar como:

$$F_1(t) = \int_0^t S(s) \cdot h_1(s) \cdot ds$$

Donde $S(s) = e^{-H_1(s)-H_2(s)}$ es la función de supervivencia en el tiempo s, determinada tanto por el evento de interés como por el competitivo.
Los modelos de riesgo de causa específica (llamados Cox-CR o Cox-CSH) son más adecuados para estudiar la etiología de las enfermedades (el riesgo que cada covariable plantea de forma independiente), mientras que el modelo de Fine-Gray se utiliza para predecir el riesgo (función de incidencia o CIF) de un evento de un individuo.

4. ANÁLISIS DE DATOS

Como en todos los modelos de supervivencia (o tiempo hasta evento), necesitamos una variable censora (que recoge el evento) y un tiempo de supervivencia de cada individuo hasta el evento de interés.
Lo primero que debemos hacer es construír una variable censora que recoja, del 0 (censura sin evento) al N siendo N el último evento de interés (por ejemplo muerte). El tiempo de supervivencia de cada individuo será el tiempo hasta el primero de los eventos que sucedan, aún en el caso de que experimente varios, o fallezca en un corto periodo tras el evento.
A continuación seleccionaremos una base de datos sólo con la variable censora, la variable supervivencia y el resto de covariables que queremos incluír (p.ej edad, sexo), pues deberemos eliminar todas las filas con valores faltantes.
Las librerías de R `survival`, `cmprsk` y `riskRegression` nos permitirán construír los modelos.

5. CONCLUSIONES

El análisis de riesgos competitivos permite reducir el sesgo de los estimadores de riesgo de un evento en presencia de un evento competitivo.

Referencias

[1] L. Teixeira, A. Rodrigues, M.J. Carvalho, A. Cabrita, D. Mendonça, Modelling competing risks in nephrology research: An example in peritoneal dialysis, *BMC Nephrology* 14 (2013) 110. doi:10.1186/1471-2369-14-110.

[2] J.P. Fine, R.J. Gray, A Proportional Hazards Model for the Subdistribution of a Competing Risk, *Journal of the American Statistical Association.* 94 (1999) 496–509. doi:10.1080/01621459.1999.10474144.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

ESTIMACIÓN NON PARAMÉTRICA DA DENSIDADE NO PLANO CON R

María Bugallo Porto[1]

[1]Estudante do Máster en Técnicas Estatísticas. Universidade de Santiago de Compostela.

RESUMO

Dado un conxunto de datos reais, unha tarefa estatística de interese é obter conclusións sobre como é o seu comportamento co fin de facer predicións, estimacións,... Intuitivamente resulta de gran utilidade estudar os posibles valores que pode tomar unha certa variable ou vector aleatorio e a probabilidade de ocorrencia dos mesmos, chegando así ao concepto de distribución de probabilidade dunha variable ou vector aleatorio. Calquera cálculo ou proceso que os involucre, como pode ser a obtención da súa media, varianza, función característica,..., os contrastes de hipóteses, os modelos de regresión,... involucran á distribución probabilística dos mesmos.

A partir dunha mostra e co fin de estudar a concentración da poboación á que pertence, o máis habitual é que se empreguen a función de distribución e a función de densidade, ambos conceptos moi estreitamente relacionados. Polo tanto, un problema estatístico fundamental é a estimación da función de densidade dunha variable ou vector aleatorio a partir da información proporcionada por unha mostra do mesmo.

En ocasións coñecemos información externa á mostra que condiciona a estimación da función de densidade, restrinxíndoa a certa clase de funcións. Nese caso, un posible enfoque é a estimación paramétrica da función de densidade, que consiste en supoñer que a función a estimar pertence a algunha familia paramétrica de funcións coñecidas e estimar os parámetros descoñecidos de tales distribucións a partir dos datos da mostra. Sen embargo, en moitos casos, o enfoque paramétrico non ten sentido porque, ou ben non coñecemos información externa á mostra, ou ben existen dúbidas sobre a validez desta información.

Como alternativa podemos non impoñer ningún modelo paramétrico fixo a función de densidade, permitindo que a función a estimar adopte case calquera forma posible, sen máis que esixir que sexa unha densidade. Chegamos así a estimación non paramétrica da función de densidade, que en cada punto considera os datos da mostra similares para realizar a estimación e que pode ter en conta ou non a proximidade. Os estimadores non paramétricos da densidade máis empregados son o estimador histograma e o estimador tipo núcleo.

Claramente a dimensión dunha mostra pode ser calquera número enteiro positivo, pero se nos restrinximos unicamente ás mostras bidimensionais, estamos considerando conxuntos de datos entre os que se atopan, por exemplo, aqueles que engloban posicións nun mapa (empregaremos como exemplo ilustrativo as posicións dos niños de avespa velutina en Galicia entre os anos 2016, 2017 e 2018, cuxos contornos de nivel da estimación da densidade tipo núcleo se representan na Figura 1). Fixándonos no caso bidimensional, o entorno e linguaxe de programación estatística R Core Team (2020) (software libre) ten implementado o estimador histograma e o estimador tipo núcleo. A finalidade desta charla é amosar como R permite obter estimadores non paramétricos

da densidade no plano e visualizar diversas gráficas que nos faciliten a interpretación dos datos, así como nos permitan extraer conclusión sobre o seu comportamento. Para solventar o primeiro dos obxectivos introduciremos a librería **KernSmooth** e para o segundo deles, é dicir, para a representación gráfica, a librería **plot3D**.

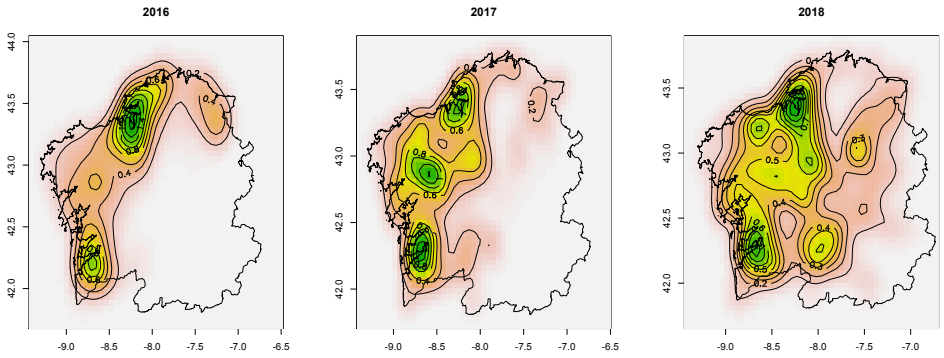

Figura 1: Evolución temporal da estimación da densidade dos niños de avespa velutina en Galicia entre 2016 e 2018, ambos inclusive, empregando curvas de nivel e mapas de calor. A escala de cores representa o aumento da concentración dos niños, dende branco ata tonos verdes.

Palabras e frases chave: Estimación da densidade, estimador histograma, estimador tipo núcleo, conxunto de nivel.

Referencias

[1] Chacón, J. E. y Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications.* Chapman and Hall.

[2] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[3] Wand and Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall, London, 1ª edition.

[4] Karline Soetaert (2019). R package plot3D: Plotting Multi-Dimensional Data (version 1.3). URL https://cran.r-project.org/web/packages/plot3D/.

[5] Matt Wand (2020). R package KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995) (version 2.23-17). URL https://cran.r-project.org/web/packages/KernSmooth/.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

A NEW R-PACKAGE FOR THE ANALYSIS OF THE FISHERIES POPULATION UNDER UNCERTAINTY

M. Cousido-Rocha[1], S. Cerviño[1], M.G. Pennino[1]

[1]Instituto Español de Oceanografía (IEO), Vigo

ABSTRACT

Rfishpop is an R package for analyzing exploited populations. More precisely, our package implements a completed Management Strategy Evaluation (MSE) cycle. MSE is a tool that scientists can use to simulate the behaviour of a fisheries system and allow them to test whether potential management procedures can achieve pre-agreed management objectives. Here, we describe all the interlinked model structures in MSE and its implementation. Furthermore we provide the main conclusions and a discussion about open issues.

Keywords: Management strategy evaluation, management procedures, operating model, assessment model.

1. INTRODUCTION

The analysis of the dynamic of a population has become a fundamental tool in ecology, conservation biology, and particularly in fisheries science to assess the status of exploited resources. Uncertainty is an inherent component in fishery systems that makes difficult taking management decisions. Here, we present Rfishpop (available on https://github.com/IMPRESSPROJECT/Rfishpop) a package to deal with with uncertainty for analyzing exploited populations in R. More precisely, Rfishpop package address such aims implementing a completed Management Strategy Evaluation (MSE) cycle (Punt et al. 2016 and Kell et al., 2007), a simulated approach explicitly designed to identify fishery rebuilding strategies and ongoing harvest strategies that are robust to uncertainty and natural variation.

2. MSE METHODOLOGY

In this section we provide a brief but complete description of MSE methodology.
A prototypical MSE incorporates a number of interlinked model structures, a schematic MSE model. The steps for a MSE cycle, as schematically showed in Figure 1, can be described as follows.

1. **Population dynamics and fishing activity (Operating Model, OM)**: An operating model is typically used to generate "true" ecosystem dynamics including the natural variations in the system.

2. **Data collection**: Data are sampled from the OM to mimic collection of fishery dependent data and research surveys (and their inherent variability).

3. **Data analysis, stock assessment and Harvest Control Rule (HCR)**: These data are passed to the assessment model. Based on this assessment and the HCR, a management action is determined (e.g., a change in the Total Allowable Catches, TAC).

4. **Implementation of the HCR**: Corresponding fleet effort and catch are then modelled, potentially allowing for error in implementation, and resulting catches are fed back into the operating model, OM.

By repeating this cycle the full management process is modelled. It is possible to test the effect of modifying any part of this cycle including changes to the operating model, assumptions about noise, etc. Alternative Management Procedures (MPs) can be compared by running many stochastic simulations, each for several years, to identify the performance of a rule according to different metrics under the likely range of conditions.

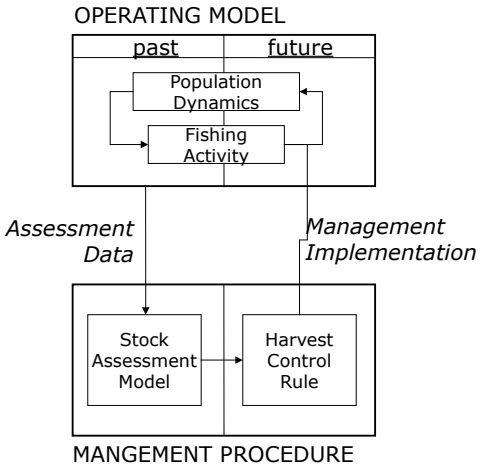

Figure 1: Schematic of a management strategy evaluation model.

3. IMPLEMENTATION

After describing a prototypical MSE cycle in this section we provide some details about how this methodology has been implemented in our R package.
In its current state, the package includes tools to simulate the real dynamics of a fishery using a generic age-structured operating model (see Chapter 12 of Haddon, 2011). The OM models a biological system with recruitment, growth, maturity and natural mortality and a fishery system were fishing intensity and selection. This allows to implement structural uncertainty having different options for each process and natural stochasticity playing with variability in these processes.
Once the exploited population has been generated through the OM, the package also contains a set of methods to estimate biological reference points as Maximum Sustainable Yield (MSY) reference points, see Chapter 8 of Hart and Reynolds (2002). These points allow to identify management targets in terms of fishing intensity, population status and yield.
The package also contains statistical methods for sampling data from the OM simulating sampling error, which is another source of uncertainty in fishery management. These methods provides different data types which can suit different assessment methods, from simple data-limited methods to more complex age or length-structured methods (examples of assessment models can be found in Chapters 6 and 7 of Haddon, 2002).
As we mentioned above, the data obtaining from the sample functions are passed to the assessment model. Our package does not develop any new assessment model as the idea is to implement the already existents ones. The package contains specific functions to change the format of the data reported by Rfishpop into the required format of the assessment model function.

Finally, the package contains functions to implement the resulting management action, determined from the assessment and the HCR, projecting our exploited population through the years on based of catches or effort established by the management action.

4. CONCLUSIONS

The described functions of Rfishpop package allow to verify the performance of management strategies or procedures in different settings generated from the OM. The package is also useful to check the performance of assessment models when some their assumptions are violated or some parameters are misspecificated.
It is important to stand out that this package is an open project, future aims focus on introducing new posibilities at some steps of the MSE cycle and also improvements in some of the procedures already implemented.

5. ISSUES

As we have explained above, MSE cycle contains a number of interlinked model structures which are not simple, and furthermore this cycle is not run once, we need to run the cycle over and over, once the resulting catches of the managament action are fed back intonthe operating model, OM. R allows to implement a complicated procedure as the MSE methodology. However, the problem is as often the time of computation required by R to run all the process. Of example, at Step 4 of the cycle we need to find the effort corresponding to the catches derived from the management action. We need to do that for each of the years through the population must be projected, and furthermore for each iteration of the population generated from OM, due to the uncertainty introduced by the OM we have a large number of stochastic populations. Hence, if we use the function "optimize" to find such effort it is too slow, even using "parLapply" function to parallelize the code for each stochastic iterations. We will discuss ideas about how to solve these issues.
Other problems relate to statistical procedures, for example definition of stochastic matrices. For the length age structured matrix we need to introduce the posibility to define stochastic matrices from the deterministic one (L). However, if we generate the value of the stochastic matrix for each row i (age) and column j (year) from a normal distribution which mean L_{ij} (the corresponding deterministic value) and variance derived from some coefficient of variation a problem appears. The resulting matrix (L^s) verifies $L^s_{ij} > L^s_{i+1,j+1}$ which is not possible since the fish never reduced its length. For example, the fish at age 4 and year 2019 has a length 10cm whereas at year 2020 and age 5 years its length is 8 cm.
We will discuss ideas about how to solve the above issues. Other problems are also related to the required time of computation and to the introduction of uncertainty through stochastic iterations.

6. FURTHER INFORMATION

Github repository: https://github.com/IMPRESSPROJECT/Rfishpop.
Tutorials: https://github.com/IMPRESSPROJECT/Tutorials-Rfishpop.
Brief explanation: Video (Youtube).

ACKNOWLEDGMENTS

The authors thank the financial support of the project IMPRESS (RTI2018-099868-B-I00) project, ERDF, Ministry of Science, Innovation and Universities - State Research Agency, and also of GAIN (Xunta de Galicia), GRC MERVEX (nº IN607-A 2018-4).

References

[1] Haddon, M. (2011). Modelling and Quantitative Methods in Fisheries. Chapman and Hall, U.S.A.

[2] Hart, P.J.B, and Reynolds, J.D. (2002). Handbook of Fish Biology and Fisheries: Fisheries, Volume 2. Blackwell Science Ltd, U.S.A

[3] Kell, L., Mosqueira, I., Grosjean, P., Fromentin, J., Garcia, D., Hillary, R., Jardim, E., Mardle, S., Pastoors, M., Poos, J. et al. (2007). FLR: an open-source framework for the evaluation and development of management strategies. ICES Journal of Marine Science, 64(4), 640:645.

[4] Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A., and Haddon, M. (2016). Management strategy evaluation: best practices. Fish and Fisheries, 17(2), 303-334.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

O EMPREGO DE R NA DETECCIÓN DAS CARACTERÍSTICAS MÁIS INFLUENTES NA CLASIFICACIÓN DE PACIENTES INFECTADOS POR COVID-19 EN GALICIA

Laura Davila-Pena[1], Balbina Casas-Méndez[1] e Ignacio García-Jurado[2]

[1]IMAT, Instituto de Matemáticas, MODESTYA, Grupo de Modelos de Optimización, Decisión, Estatística e Aplicacións, Departamento de Estatística, Análise Matemática e Optimización, Facultade de Matemáticas, Universidade de Santiago de Compostela.
[2]CITIC, Centro de Investigación en Tecnoloxías da Información e das Comunicacións, MODES, Grupo de Modelización, Optimización e Inferencia Estatística, Departamento de Matemáticas, Facultade de Informática, Universidade da Coruña.

RESUMO

Un problema de clasificación consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos. Ditas predicións baséanse no coñecemento obtido a través dunha mostra de individuos para os que sabemos os valores dos atributos e da variable resposta. Os problemas de clasificación pódense abordar dende as técnicas do *machine learning*. Na literatura do *machine learning*, propuxéronse e analizáronse numerosos clasificadores (ver, por exemplo, [4]). Moitos deles, ademais de clasificar, permítennos avaliar a importancia que os diversos atributos tiveron na clasificación dun individuo concreto. En [5] introdúcese un procedemento xeral (é dicir, válido para calquera clasificador) para avaliar dita importancia. Este procedemento baséase no valor de Shapley [6] para xogos cooperativos.

Nesta charla analizamos unha base de datos de pacientes de Galicia infectados por COVID-19 dende o 06/03/2020 ata o 07/05/2020. O noso obxectivo é estudar a influencia de varias características dos pacientes en tres variables resposta binarias de especial interese: a necesidade de hospitalización, a necesidade de ingreso en UCI e o falecemento. A nosa énfase non está na clasificación preditiva de novos pacientes, senón na análise das características que influíron en que os pacientes cuxa historia completa coñecemos tiveran unha resposta positiva nas variables binarias indicadas.

Este estudo lévase a cabo mediante `R`, e o clasificador empregado para a análise é o clasificador *random forest* [3], implementado na libraría `RWeka` [1], interface en `R` de Weka [2].

Palabras e frases chave: machine learning, clasificación, RWeka, influencia de atributos, valor de Shapley, covid-19.

Referencias

[1] Hornik, K., Buchta, C. e Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, 24, 225–232.

[2] Witten, I.H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition*. Morgan Kaufmann, San Francisco.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

[4] Fernández-Delgado, M., Cernadas, E., Barro, S. e Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.

[5] Strumbelj, E. e Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 1–18.

[6] Shapley, L.S. (1953). A value for n-person games. In Kuhn, H.W. e Tucker, A.W. (eds.): *Contributions to the Theory of Games (AM-28)*, Volume II, 307–318. Princeton University Press.

Emprego de R na predición cooperativa de variables relacionadas coa pandemia do Covid-19

Rubén Fernández Casal[1], Carlos Fernández-Lozano[2] e José A. Vilar Fernández[1]

[1] CITIC, Grupo MODES, Departamento de Matemáticas, Universidade da Coruña
[2] CITIC, Grupo RNASA-IMEDIR, Departamento de Ciencias da Computación e Tecnoloxías da Información, Universidade da Coruña

RESUMO

O proxecto "Predición Cooperativa" xurdiu dentro da iniciativa "Matemáticas contra o coronavirus" promovida polo Comité Español de Matemáticas (CEMat). Como resultado desenvolveuse unha web interactiva empregando R (https://covid19.citic.udc.es) para monitorizar e predicir a curto prazo variables relevantes na propagación do Covid-19. Nesta web proporcionáronse "predicións cooperativas" (meta-predicións), a horizontes de 1 a 7 días para cada comunidade autónoma e variable de interese, combinando predicións baseadas en distintos métodos que proporcionaban regularmente un gran número de grupos de investigación de xeito independente.

Palabras e frases chave: meta-predictores, series de tempo, visualización interactiva.

1. O PROXECTO PREDICIÓN COOPERATIVA

O proxecto "Predición Cooperativa" xurdiu no marco do programa "Matemáticas contra o coronavirus" (http://matematicas.uclm.es/cemat/covid19), posto en marcha polo CEMat, co fin de aproveitar as capacidades de análise e modelado da comunidade española de matemáticos e estatísticos para axudar a comprender e xestionar a crise sanitaria causada pola epidemia Covid-19. O obxectivo principal era a construción de meta-predictores para proporcionar ás autoridades información sobre o comportamento a curto prazo de variables de gran interese na propagación do virus Covid-19. Nun contexto de alta incerteza, como ocorría neste caso especialmente ao principio da pandemia, a combinación de preditores pode axudar a evitar os problemas de mala especificación dos preditores individuais, obtendo predicións máis estables e aumentando a precisión (e.g. Armstrong, 2001). Para mais detalles sobre este proxecto ver Vilar-Fernández *et al.* (2020).

Consideráronse as seguintes variables: ingresos en UCI, falecidos, hospitalizados, casos confirmados e novos contaxios. Os datos obtíñanse dos publicados diariamente polo Instituto de Salud Carlos III (https://cnecovid.isciii.es/covid19; tamén dispoñibles en https://rubenfcasal.github.io/COVID-19 no formato da iniciativa). Esta información, ademais das predicións cooperativas, podía ser consultada na web do proxecto (https://covid19.citic.udc.es) de xeito interactivo. Esta aplicación desenvolveuse en R empregando *shiny* (Chang *et al.*, 2020) e os paquetes *leaflet* (Cheng *et al.*, 2019), *ggplot2* (Wickham, 2016) e *ploty* (Sievert, 2020), entre outros. O despregamento realizouse nun contedor docker cun orquestrador swarm encargado de equilibrar a carga para evitar a saturación debida a un elevado número de accesos simultáneos (o pico máximo de visitas nun día foi de 1500).

O 1 de abril fíxose un chamamento a todos aqueles investigadores dispostos a desenvolver modelos para predicir a evolución da pandemia solicitando a súa colaboración para que enviasen diariamente as súas predicións. A resposta foi moi positiva, inscribíndose 62 grupos de investigación (máis de 130 investigadores en total), dos cales 49 participaron activamente. Unha lista non exhaustiva está dispoñible na pestana "Predición cooperativa: información/Investigadores colaboradores" da web do proxecto.

Os participantes proporcionaron predicións de polo menos unha das comunidades autónomas ou de toda España, de a lo menos unha das variables de interese e para algúns dos horizontes de 1 a 7 días (non necesariamente do rango completo), que tiñan que enviar diariamente nun ficheiro de excel antes das 19h. Antes de rematar o día calculábanse as predicións cooperativas e publicábanse na web.

Desenvolveuse código en R para automatizar todo o proceso de xestión de datos, cálculo de predicións combinadas e erros de predición e xeración de informes, empregando os paquetes *dplyr* (Wickham *et al.*, 2020) e *rmarkdown* (Allaire *et al.* , 2020), entre outros. O código é robusto para tratar as diferentes combinacións de predicións e contribucións distribuídas irregularmente ao longo do tempo (nun primeiro momento intentouse empregar ferramentas xa dispoñibles, como o paquete *ForeComb*; Weiss *et al.*, 2018). O código principal está dispoñible en aberto no seguinte enlace:
https://github.com/rubenfcasal/COVID-19/tree/master/prediccion_cooperativa
(un compromiso de confidencialidade impide proporcionar as predicións individuais dos grupos participantes, intentaremos que no futuro estean en aberto mantendo o anonimato).

Incluíronse diferentes métodos de combinación, desde métodos sinxelos ata métodos máis sofisticados, con pesos estimados segundo criterios de optimalidade, que requirían dun adestramento (en cada combinación de variable de interese, rexión administrativa e horizonte). Na pestana "Predición cooperativa: Información/Ficha técnica" da web do proxecto móstranse mais detalles sobre os distintos métodos (ver tamén Vilar-Fernández *et al.*, 2020). As predicións cooperativas, xunto con medidas das súas precisións, podían ser consultadas na pestana "Predición cooperativa: Resultados". Para cada combinación de variable e CCAA o usuario podía interactuar con gráficos e táboas dinámicas.

Os múltiples cambios e problemas cos datos oficiais proporcionados polo ISCIII dificultaron notablemente a avaliación da precisión das predicións (ademais de ter que refacer o código continuamente e desanimar ós colaboradores a seguir participando). Houbo todo tipo de problemas, incluíndo períodos sen información oficial dispoñible e numerosas rectificacións das series. Isto obrigou a manter un histórico dos valores reportados e crear un filtro para ter en conta todos estes cambios (cando se producía un cambio invalidábanse as predicións a horizontes futuros). Pódese consultar un informe no seguinte enlace:
https://rubenfcasal.github.io/COVID-19/acumula2_hist/Informe_acumula2_hist.html,
que pode servir tamén para ver ós problemas dos datos reportados polas distintas CCAA. Entre os distintos problemas houbo períodos sen información dispoñible e numerosas rectificacións dos valores anteriormente reportados.

As predicións cooperativas comezaron a calcularse o 2 de abril (dende o 9 de abril as que requirían dun adestramento) de xeito diario, salvo os períodos sen información oficial dispoñible ou cambios nas series de datos. Debido a enorme carga de traballo, dende o 18 de maio as predicións cooperativas pasan a realizarse tres días a semana

(luns, mércores e venres). O 26 de maio actualizouse por última vez a serie de datos do ISCIII o que obrigou a paralizar a iniciativa.

Durante a vixencia do proxecto os colaboradores aportaron 299108 predicións individuais e calculáronse 209160 predicións cooperativas. Actualmente estase a desenvolver unha aplicación Shiny que permita examinar e comparar gráfica e analiticamente o comportamento de todos os preditores (combinados e individuais, mantendo o anonimato) no escenario elixido (variable, CCAA, horizonte). Hai que ter en conta que resulta moi complicado analizar as precisións, e especialmente obter conclusións globais, debido ós cambios e incoherencias nos datos oficiais, ás diferencias no número e na distribución no tempo das predicións individuais e á gran cantidade de escenarios (700 en total).

O proxecto recibirá financiamento do FONDO SUPERA COVID-19 tras resultar seleccionado na Convocatoria Crue-CSIC-Santander, o que permitirá a contratación durante un ano de dous investigadores para continuar co desenvolvemento das ferramentas. O obxectivo sería melloralas para facilitar o seu uso no futuro, incluíndo a posibilidade de empregalas para outras variables de potencial interese e noutros ámbitos de estudio (por exemplo en áreas sanitarias ou noutros países).

2. CONCLUSIÓNS

Resulta fundamental mellorar a calidade das series de datos proporcionadas polos gobernos autónomos. Para facilitar e automatizar a adquisición de datos, é necesario unificar a definición de variables. Habería que esixir ás autoridades competentes que se poida dispoñer de series de datos fiables e homoxéneas en aberto.

Os resultados apoian o interese en combinar predicións, obtéñense predicións máis estables e en xeral mais precisas: "a previsión combinada pode ser mellor que a mellor pero non peor que a media" (Armstrong, 2001). Observouse tamén o denominado “forecast combination puzzle”: preditores combinados simples como a media ou a mediana son altamente competitivos e a miúdo máis precisos que métodos mais sofisticados construídos mediante adestramento.

AGRADECEMENTOS

Os autores queren dar as grazas a tódolos grupos de investigación e investigadores que colaboraron enviando as súas predicións, xa que sen o seu tempo e dedicación desinteresada este proxecto non sería viable.

Moitas grazas tamén ós demais membros do equipo de traballo de predición cooperativa: Ricardo Cao Abad (UDC), Daniel Barreiro Ures (UDC), Ana Almécija Pereda (UDC), Alfonso Gordaliza Ramos (UVA), Luis Ángel García Escudero (UVA) e Pablo Montero Manso (Monash University, Australia) polo seu esforzo e dedicación desinteresada.

Agradecer tamén ó Comité Español de Matemáticas (CEMat) o seu apoio e interese no proxecto, e ó Centro de Investigación en Tecnoloxías da Información e a Comunicación (CITIC) da Universidade de Coruña, pola asistencia prestada e por proporcionar a infraestrutura necesaria.

Referencias

[1] Allaire, J.J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. y Iannone, R. (2020). rmarkdown: Dynamic Documents for R. R package version 2.1. https://rmarkdown.rstudio.com.

[2] Armstrong, J.S. (2001). Combining forecasts. En: Principles of Forecasting: A Handbook for Researchers and Practitioners (Cap. 4), 417-439, Kluwer Academic Publishing.

[3] Chang, W., Cheng, J., Allaire, J.J., Xie, Y. y McPherson, J. (2020). shiny: Web Application Framework for R. R package version 1.4.0.2. https://shiny.rstudio.com.

[4] Cheng, J., Karambelkar, B. y Xie, Y. (2019). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3. http://rstudio.github.io/leaflet.

[5] Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall. https://plotly-r.com.

[6] Vilar-Fernández, J.A., Fernández-Casal, R. y Fernandez-Lozano, C. (2020). Covid-19 projections for Spain using forecast combinations. BEIO, 36 (2), 99-125. http://www.seio.es/BBEIO/BEIOVol36Num2/files/assets/common/downloads/publication.pdf#page=15.

[7] Weiss, Ch. E., Raviv, E., y Roetzer, G. (2018). Forecast combinations in R using the ForecastComb package. The R Journal, 10(2), 262-281.

[8] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. http://ggplot2.tidyverse.org.

[9] Wickham, H., François, R., Henry, L. y Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://dplyr.tidyverse.org/index.html.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

DESARROLLO DE UN ATLAS DE MORTALIDAD EN CASTILLA-LA MANCHA CON R

Virgilio Gómez Rubio[1] y Francisco Palmí Perales[1]

[1]Departamento de Matemáticas, E.T.S. Ingenieros Industriales - Albacete, Universidad de Castilla-La Mancha

RESUMEN

El desarrollo de atlas de mortalidad en áreas pequeñas es importante en salud pública puesto que da información sobre la variación espacial y temporal del riesgo de fallecer por determinadas causas de mortalidad. La elaboración de un atlas de mortalidad requiere el análisis de datos de población y mortalidad, así como de información sobre las áreas geográficas de la región de estudio. En este trabajo describimos cómo se ha desarrollado un atlas de mortalidad en Castilla-La Mancha a nivel munícipal en el período 2003-2014. Para ello hemos empleado el software estadístico `R` y algunos paquetes, así como *Shiny server* para crear una versión interactiva del atlas.

Palabras y frases clave: epidemiología, estadística espacial, inferencia Bayesiana, mortalidad, R

1. INTRODUCCIÓN

Las autoridades sanitarias recogen de manera regular datos sobre la mortalidad y morbilidad de la población. Esta información se suele utilizar de distintas maneras para la gestión y toma de decisiones sanitarias. Una opción es el desarrolllo de atlas de mortalidad, para estudiar la distribución de la mortalidad por distintas causas en áreas pequeñas, como municipios o zonas básicas de salud. A continuación vamos a describir la realización de un atlas de mortalidad a nivel municipal en Castilla-La Mancha durante el período 2003-2014 [2].

2. DATOS

Los datos de mortalidad a nivel individual fueron adquiridos del Instituto Nacional de Estadística dentro del marco de un convenio de investigación (ver Agradecimientos). Los datos de población se obtuvieron también del INE. Asimismo, se han utilizado las poligonales de los municipios de Castilla-La Mancha disponibles en la web del INE.

3. MODELIZACIÓN

Para la gestión de los datos y la modelización de los mismos se ha utilizado el software estadístico R [4]. En primer lugar, se han obtenidos las tasas específicas por edad y sexo para,aplicándolas a la estructura de población de los municpios, calcular el número de casos esperado en cada municipio de cada una de las causas estudiadas. A partir de aquí, se calcula la razón de mortalidad estandarizada (RME) como primer indciador del riesgo relativo por municipio. El RME está definido como el cociente entre los caos observados y esperados, de manera que valores mayores de 1 indican un exceso de mortalidad en el municipio. También se calcula el RME por año para poder estudiar la variación temporal de la mortalidad.
Como el RME es muy intestable en municipios poco poblados, se han calculado estimadores del riesgo relativo (RR) suavizado usando los modelos de Besag, York y Mollié [1] usando modelos

espaciales, y también los modelos de Knorr-Held [3] usando modelos espacio-temporales. El ajuste de modelos se ha hecho utilizando el paquete `INLA` [5].
De esta manera, obtenemos estimaciones del RR suvizado a nivel municipal para todo el período de estudio con los modelos espaciales, y estimaciones del RR suavizado por muncipio y año con los modelos espacio-temporales.

4. DESARROLLO DEL ATLAS

Una vez que se han obtenido las estimaciones del RR (RME y valores suavizados), el siguiente paso es construir el atlas. Para ello, se ha incluido en primer lugar una introducción con la metodología utilizada. También se ha incluido un estudio de la población a nivel municipal durante todo el período puesto que la mayoría de las municipios de Castilla-La Mancha pertenecen a la llamada "España vaciada". Un ejemplo de las tasas suavizadas calculadas para todo el período de estudio puede verse en la Figura 1.

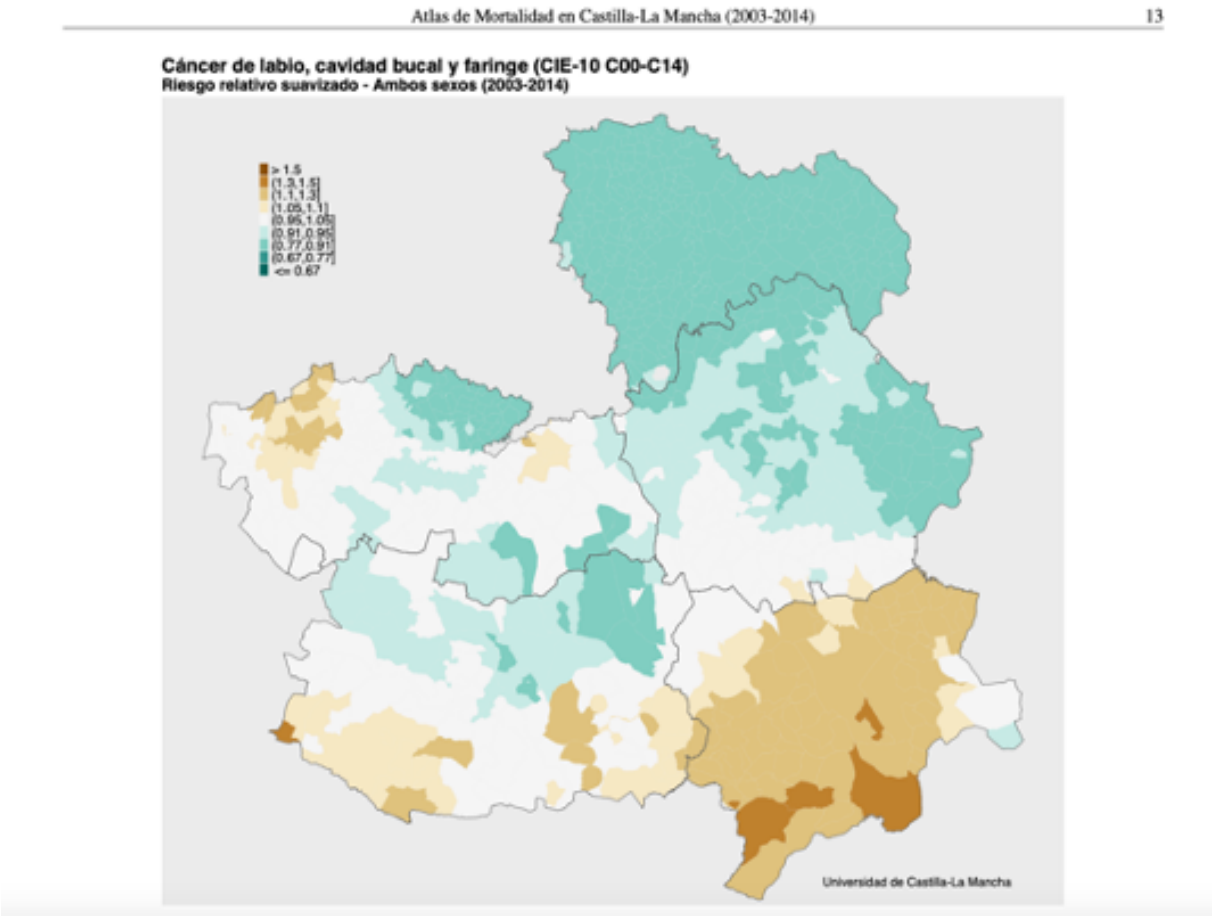

Figura 1: Riesgos relativos suavizados para ambos sexos para el período 2003-2014 calculados usando modelso espaciales (cáncer de labio, cavidad bucal y laringe).

El libro del atlas, para ser impreso, se ha desarollado con el paquete `bookdown` [7]. Una versión on-line del mismo se ha desarrollado también usando una App de shiny [6]. Ambos están disponibles en la web http://atlasmortalidad.uclm.es . La Figura 2 muestra la portada del atlas on-line.

5. CONCLUSIONES

`R` es un software estadístico muy potente para el análisis de datos. En este trabajo hemos mostrado que es posible utilizarlo para el análisis de datos de epidemiología y salud pública. En concetro, para el desarrollo de un atlas de mortalidad en Castilla-La Mancha.

Figura 2: Portada de la versión on-line del atlas.

AGRADECIMIENTOS

El desarrollo del atlas ha sido posible gracias al plan propio de investigación de la Universidad de Castilla-La Mancha, los proyectos PPIC-2014-001-P y SBPLY/17/180501/000491 financiados por la Consejería de Educación, Cultura y Deportes de la Junta de Comunidades de Castilla-La Mancha y FEDER, el proyecto MTM2016-77501-P del Ministerio de Economía y Competitividad, el proyecto PID2019- 106341GB-I00 del Ministerio de Ciencia e Innovación. Además, Francisco Palmí Perales ha contado con un contrado doctoral del plan propio de la Universidad de Castilla-La Mancha para el desarrollo de su tesis doctoral.
Los datos relativos a las defunciones han sido adquiridos al Instituto Nacional de Estadística (INE) a través del convenio de cesión de datos regulado por el: "Protocolo de cesión de ficheros finales de microdatos de defunciones, según la causa de muerte, del Instituto Nacional de Estadística a la Universidad de Castilla-La Mancha para el desarrollo del proyecto Modelos Jerárquicos Bayesianos en Medio Ambiente, Oncología y Salud".

Referencias

[1] Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43, 1–20 (1991). https://doi.org/10.1007/BF00116466

[2] Gómez Rubio, V., y Palmí Perales, F., eds. (2020). *Atlas de mortalidad en Castilla-La Mancha 2003-2014*. Editorial Bomarzo.

[3] Knorr-Held, L. (2000), Bayesian modelling of inseparable space-time variation in disease risk. Statist. Med., 19: 2555-2567.

[4] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[5] Rue, H., Martino, S. and Chopin, N. (2009), Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion), Journal of the Royal Statistical Society B, 71, 319-392.

[6] Chang, W., Cheng, J., Allaire, J.J., Xie, Y. and McPherson, J. (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny

[7] Xie, Y. (2016). bookdown: Authoring Books and Technical Documents with R Markdown. Chapman and Hall/CRC. ISBN 978-1138700109

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

PROJECTMANAGEMENT: UN PAQUETE DE R PARA A XESTIÓN DE PROXECTOS

Juan Carlos Gonçalves-Dosantos[1], Ignacio García-Jurado[1] e Julián Costa[2]

[1]Grupo MODES, CITIC e Departamento de Matemáticas, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain
[2]Grupo MODES, Departamento de Matemáticas, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain

RESUMO

A xestión de proxectos é un importante corpo de coñecementos e prácticas que comprenden a planificación, organización e control dos recursos para lograr un ou máis obxectivos predeterminados. Presentamos ProjectManagement, un novo paquete de R que proporciona as ferramentas necesarias para a xestión de proxectos: planificación de un calendario, nivelación de recursos e reparto de custos por demora. Todas estas funcións abórdanse dende dous contextos diferentes, considerando duracións deterministas e estocásticas para as distintas actividades.

Palabras e frases chave: proxectos, demoras, recursos.

AGRADECEMENTOS

Este traballo foi financiado por a subvención MINECO MTM2017-87197-C3-1-P e por a Xunta de Galicia a través de FEDER (Grupos de Referencia Competitiva ED431C-2016-015 e Centro Singular de Investigación de Galicia ED431G/01).

Referencias

[1] Gonçalves-Dosantos J.C., García-Jurado I., Costa J. (2020). ProjectManagement: an R package for managing projects. *The R Journal* To appear.

DETERMINACIÓN DE LA CALIDAD SENSORIAL DE ALIMENTOS MEDIANTE MAPAS DE PREFERENCIA

Andrés Martínez Sánchez[1]

[1] Técnico sensorial TasteLab & SENSESBIT

RESUMEN

En el presente trabajo se realiza un análisis estadístico multivariante de los datos obtenidos tras la aplicación del análisis sensorial tanto con paneles de catadores (datos objetivos), como con consumidores (datos subjetivos).

Si bien la aproximación tradicional consiste en analizar estos datos de manera univariante, en este estudio se propone una aproximación multivariante, la elaboración de mapas de preferencia, con el objetivo de determinar la calidad sensorial de los productos de forma gráfica y con una interpretación intuitiva.

Palabras y frases clave: Análisis sensorial, mapa de preferencia, PCA, MFA.

1. INTRODUCIÓN

El análisis sensorial es un método científico que pretende medir, analizar e interpretar las respuestas de un grupo de sujetos, a los estímulos percibidos a través de los sentidos de la vista, el olfato, el gusto y el tacto, ante un producto determinado.

El termino de calidad sensorial se define con dos aspectos complementarios. El primero apela a la percepción subjetiva que tiene el consumidor de ese producto, obtenida a partir de pruebas hedónicas, y el segundo a las características intrínsecas de un producto, definidas mediante la utilización de catadores entrenados.

En el contexto de este trabajo, se busca analizar las respuestas subjetivas de los consumidores frente a determinados productos, a través de la obtención de un mapa de preferencia interno.

Para la obtención de los datos, por un lado, se les pide a un grupo de consumidores que evalúen el olor, el sabor, la textura, el aspecto y que realicen una valoración global de dichos productos, utilizando una escala de 9 puntos, correspondéndose la valoración más alta (punto 9 de la escala), con me gusta muchísimo, y la más baja (punto 1), con me disgusta muchísimo.

Por otro lado, se completa el análisis con la incorporación de las respuestas objetivas proporcionadas por un panel de catadores entrenados, que evalúan los mismos productos, tras el desarrollo de las fichas de cata correspondientes en las que se incluyen los términos que definen sus características sensoriales y las escalas de medida.

2. MAPA DE PREFERENCIA INTERNO

Un mapa de preferencia interno proporciona una representación multidimensional de productos y consumidores con el objetivo de tener una visión conjunta de la aceptación de los productos evaluados por los consumidores y llegar a establecer cuál es el producto que puede maximizar la aceptación del consumidor. Para ello, se

organiza la información proporcionada por los consumidores en la valoración global de los productos en una matriz de m filas (m = número de productos evaluados) y r columnas (r = número de consumidores que intervienen en el estudio) y se realiza un análisis de componentes principales (PCA). En este caso se hace mediante la función PCA del paquete FactoMineR [1]. Para aumentar la representatividad del PCA puede ser necesario aplicar un filtro de consumidores eliminando a los peor representados por los dos primeros componentes. A continuación, se posicionan los productos y los consumidores sobre el plano determinado por los dos primeros componentes (biplot). En este trabajo se ha optado por una representación gráfica utilizando ggplot2 [2].

De esta manera se construye lo que en análisis sensorial se denomina mapa de preferencia interno [3]. Los productos se representan como letras, y puesto que la posición de los consumidores individuales no es importante, si no la distribución de la muestra de consumidores respecto a los productos, en este trabajo se ha optado por la representación de los consumidores mediante un gráfico de densidad. A partir de la Figura 1 se puede interpretar que los productos situadas hacia el cuadrante superior derecho son los más valorados por los consumidores, y aquellos más alejados, los menos valorados.

3. MAPA COMPARATIVO DE PREFERENCIA INTERNO

La forma clásica de construír mapas de preferencia es empleando la información relativa a la valoración global de los productos. La utilización de un mapa de densidad para representar a los consumidores, ayuda a percibir la distribución de la muestra de consumidores resaltando en qué dirección aumenta la aceptación global de gran parte de los consumidores. Empleando otras variables hedónicas en la organización de los datos de partida, se podrían proporcionar nuevas distribuciones de consumidores y nuevas conclusiones sobre la calidad sensorial de los productos en estudio. Por ese motivo, se decide explorar la posibilidad de elaborar un mapa comparativo de preferencia interno (Figura 2) empleando simultáneamente toda la información de las variables hedónicas que ha sido proporcionada por los consumidores.

En la construcción del mapa comparativo de preferencia interno, en lugar de partir de PCA para posicionar los productos en el mapa, se recurre a un análisis factorial múltiple (MFA), partiendo de varias matrices de datos simultaneamente. Sobre el posicionamiento de los productos en los dos primeros factores tras la aplicación de MFA, se representan secuencialmente la densidad de los consumidores para cada variable hedónica. Con el objetivo de facilitar la interpretación y elaborar categorías de calidad sensorial se elabora un análisis clúster representado en forma de dendograma.

En este ejemplo, el enfoque tradicional univariante (test de Friedman) indica que habría diferencias estadísticamente significativas para olor, valoración global y sabor. En la Figura 2 se observa que en el único de estos aspectos en los que una parte importante de los consumidores coincide es en el atributo de olor, por lo tanto se concluye que este atributo es el principal responsable de que existan diferencias importantes en la calidad sensorial percibida por los consumidores.

4. MAPA DE PREFERENCIA INTERNO CON VARIABLES SENSORIALES SUPLEMENTARIAS

Sobre un mapa de preferencia interno se puede incorporar la información del perfil sensorial de productos, obtenido a partir de un panel de catadores entrenado.

Una posibilidad es incluir la información objetiva del panel de catadores entrenado como variables suplementarias en el PCA [4], permitiendo aproximar cuales son las características sensoriales que más definen a un producto. De esta forma, sobre el mismo gráfico se puede ver qué productos son los que tienen mayor calidad sensorial para los consumidores, y cuáles son sus características sensoriales (Figura 3).

AGRADECIMIENTOS

Este trabajo forma parte del TFM del Master de Innovación en Nutrición, Seguridad y Tecnología Alimentaria, tutorizado por Maria de las Nieves Muñoz Ferreiro y Maruxa Quiroga García, a las que me gustaría agradecer por su apoyo durante el desarrollo del mismo, así como a la empresa TasteLab por la cesión de los datos sensoriales.

Referencias

[1] Le, S., & Husson, F. (2008). *FactoMineR: A Package for Multivariate Analysis*. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

[2] *Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York*. https://ggplot2.tidyverse.org

[3] Guinard, J. X., Uotani, B., & Schlich, P. (2001). Internal and external mapping of preferences for commercial lager beers: Comparison of hedonic ratings by consumers blind versus with knowledge of brand and price. *Food Quality and Preference*, *12*(4), 243-255. https://doi.org/10.1016/S0950-3293(01)00011-8

[4] Symoneaux, R., Galmarini, M. V., & Mehinagic, E. (2012). Comment analysis of consumer's likes and dislikes as an alternative tool to preference mapping. A case study on apples. *Food Quality and Preference*, *24*(1), 59-66. https://doi.org/10.1016/j.foodqual.2011.08.013

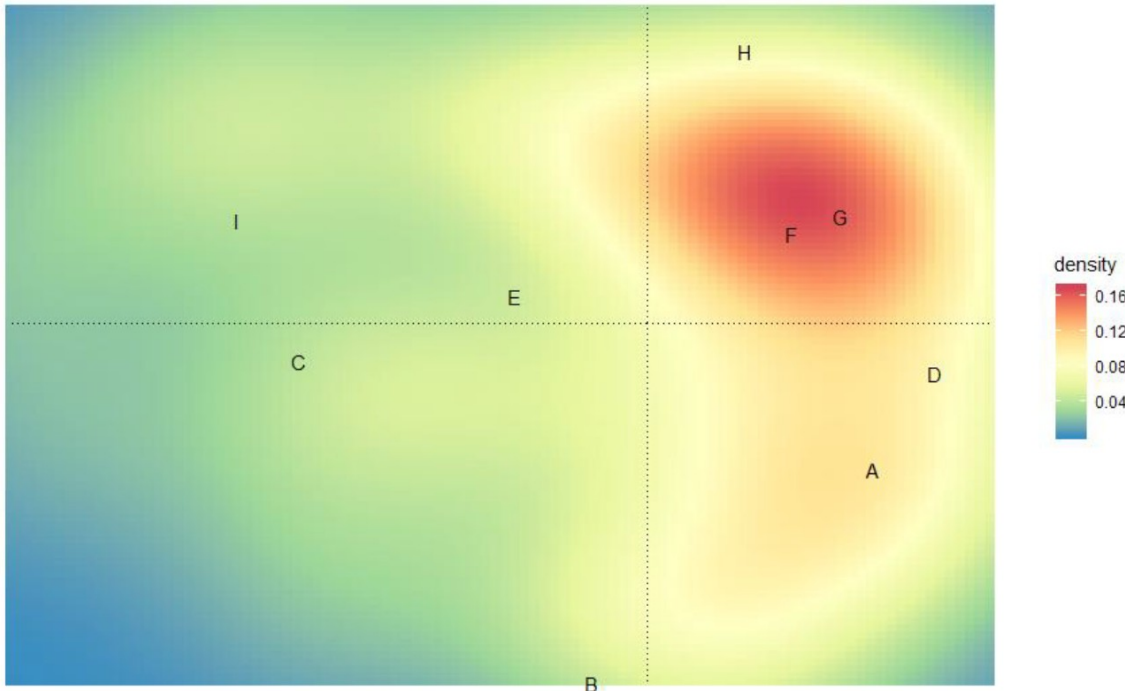

Figura 1: Mapa de preferencia interno.

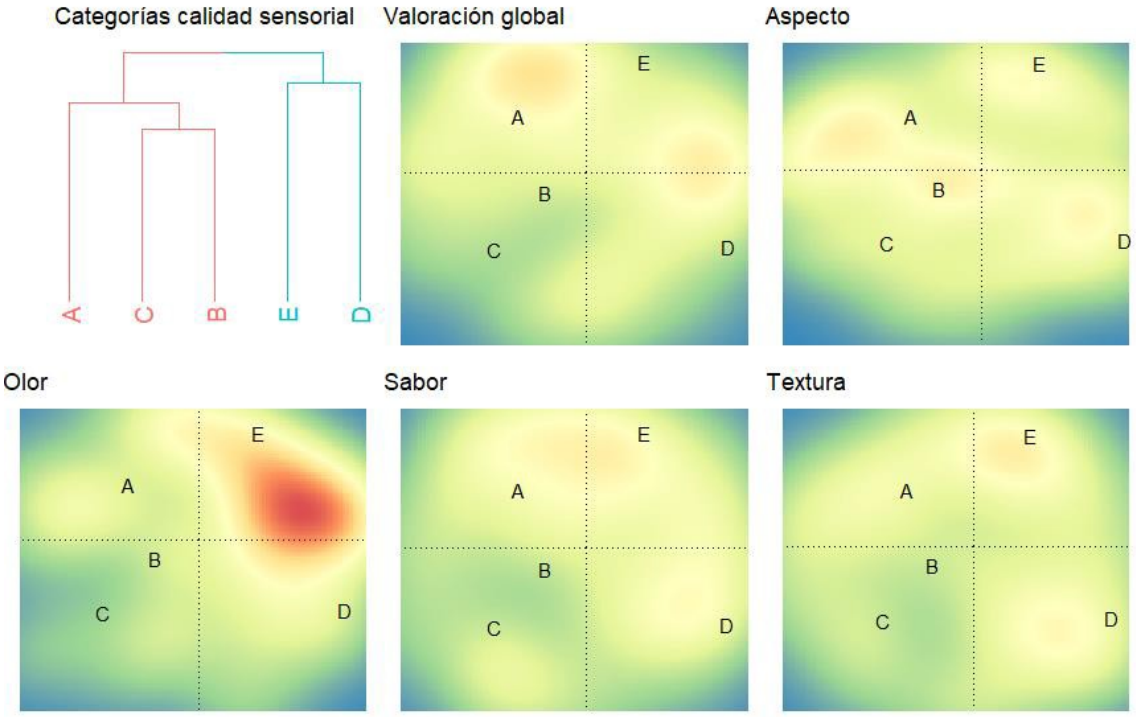

Figura 2: Mapa comparativo de preferencia interno.

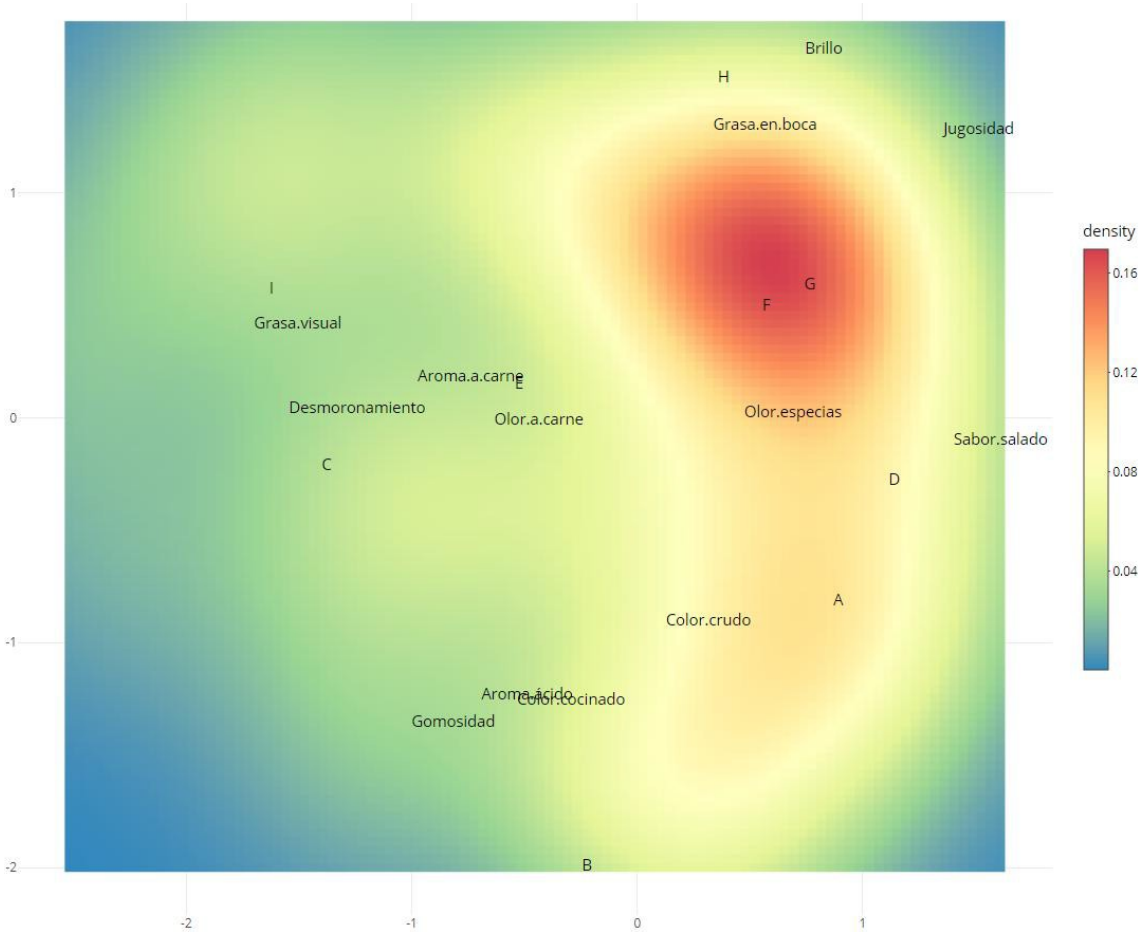

Figura 3: Mapa de preferencia interno con variables sensoriales suplementarias.

FORTLS: UN PAQUETE DE R PARA A ESTIMACIÓN DE VARIABLES DASOMÉTRICAS PARA O SEU USO EN INVENTARIO FORESTAL

Juan Alberto Molina-Valero[1], María José Ginzo Villamayor[2], Manuel Antonio Novo Pérez[3], Juan Gabriel Álvarez-González[1], Fernando Montes[4], César Pérez-Cruzado[1]

[1] Unidade de Xestión Ambiental e Forestal Sostible (UXAFORES), Departamento de Enxeñería Agroforestal, Escola Politécnica Superior de Enxeñería, Universidade de Santiago de Compostela.

[2] Departamento de Estatística, Análise Matemática y Optimización, USC de Santiago de Compostela.

[3] Instituto Tecnolóxico de Matemática Industrial (ITMATI

[4] INIA-CIFOR

RESUMO

As altas capacidades que presenta o láser escáner terrestre (TLS) para representar superficies de obxectos na contorna próxima de forma rápida, automática e cunha elevada resolución, sitúano como un dos dispositivos con maior potencial para o seu uso en Inventarios Forestais (IFs). Con todo, aínda non se estableceu como unha ferramenta operativa para este fin, principalmente debido ás dificultades que se atopan na automatización do procesado de datos. Neste traballo preséntase o paquete `FORTLS` de R, que ten como principal funcionalidade a automatización do procesado dos datos TLS para o seu uso en IFs. O paquete permite a identificación e estimación do diámetro normal das árbores e a obtención de métricas de rodal e variables dasométricas en base a un escaneo simple; o que permite reducir os tempos de toma de datos e procesado, así como aumentar o tamaño de mostra sen incrementar demasiado os custos. Estas características do paquete `FORTLS` confírenlle un alto potencial para ser utilizado en técnicas de inferencia baseadas ou asistidas por modelos. `FORTLS` permite ademais a optimización do deseño de parcelas de campo para o seu uso combinado con escaneos TLS para IFs

Palabras e frases chave: LiDAR, métodos de masa, sensores remotos, paquete R, teledetección, TLS.

AGRADECEMENTOS

A investigadora María José Ginzo agradece apoio do proxecto MTM2016-76969-P do Ministerio de Economía y Competitividad.

Os investigadores Juan Alberto Molina Valero, Juan Gabriel Álvarez González y César Pérez Cruzado agradecen apoio do proxecto AGL2016-76769-C2-2-R do Ministerio de Ciencia, Innovación y Universidades.

O investigador Juan Alberto Molina Valero agradece a concesión da axuda para a formación do profesorado universitario (FPU16/03057) do Ministerio de Ciencia, Innovación y Universidades.

VII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 15 de outubro do 2020

EL R-KWARD EN EL ANÁLISIS DE MODELOS DE REGRESIÓN CON DATOS DEL COVID-19 EN COLOMBIA

Jorge Alejandro Obando Bastidas[1], Laura Nathalia Obando[2] e María Teresa Castellanos Sanchez[3]

[1] Universidad Cooperativa de Colombia
[2] Universidad Santo Tomas
[3] Universidad de los Llanos

RESUMEN

El Covid-19 en el mundo es un fenómeno social y de salud pública que ha motivado todo proceso de investigación. La cantidad de datos e información estadística a nivel mundial, en particular Colombia, es abundante, los gráficos se han representado en forma descriptiva y algunos evidenciando una serie de tiempo, sin embargo, poco se han estudiado desde la inferencia de los modelos de regresión. El uso del software R-Kward extensión del R, permite de forma sencilla, simular estos modelos, haciendo uso de la base de datos del ministerio de salud de Colombia.

Palabras y frases clave: Modelo de regresión, R-kward, Covid-19.

1. INTRODUCIÓN

El mundo, ha percibido un alto índice de contagio del Covid-19, su expansión ha sido acelerada en diferentes países como Colombia. Todos los países han adoptado en forma particular medidas necesarias para reducir su letalidad, pero la lucha no brinda aplanamiento en la curva (1). La lectura de los datos en los medios de comunicación a todo nivel, solo permiten SU visualización, sin mostrar significancia estadística; ignorando de esta manera evaluar críticamente la evidencia publicada y mejora de decisiones complejas en la práctica diaria. (2)

Por ejemplo, proponen que en Colombia y Latinoamérica el crecimiento de contagios es exponencial, sin argumentos estadísticos y con explicaciones que propician confusión. (El país.com "Evolución del coronavirus", 7 de abril de 2020). Ante este tipo de lectura se propone en el presente artículo, los elementos de la estadística inferencial, para que, desde los supuestos lógicos de la inferencia de los modelos lineales y no lineales, se proponga una lectura real que conlleve a la explicación del comportamiento de los datos y una predicción. En (3), se observa un trabajo desarrollado en el software R-Kward, aplicación con la cual se simula el crecimiento de los datos, permite la comparación de estos modelos, y determina con el p-valor y el valor del R^2. Los modelos que evidencia esta aplicación se observan en la Tabla 1(4)

Modelo	Expresión Matemática	Coeficientes de correlación
Lineal	$y = \beta_1 + \beta_2 x$	$r_{xy} = \frac{Cov(x,y)}{S_x S_y}$
Cuadrático	$y = \beta_1 + \beta_2 x + \beta_3 x^2$	Ecuación de predicción $\hat{y} = \widehat{\beta_1} + \widehat{\beta_2} x + \widehat{\beta_3} x^2$ $r_{xy} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

Cúbico	$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$	$\hat{y} = \widehat{\beta_1} + \widehat{\beta_2} x + \widehat{\beta_3} x^2 + \widehat{\beta_4} x^3$
		$r_{xy} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

Tabla 1. Modelos matemáticos que explican una relaciona en R-Kward

Para la simulación de los modelos de regresión en el software se hará uso de los datos dispuestos en https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx del ministerio de salud colombiano. En la Tabla 2, se observa la forma como se organizan los datos.

Día de Contagio	Día Numero	Número de contagios	Acumulado
feb-27	1	142	142
feb-28	2	164	306
feb-29	3	132	438
mar-01	4	133	571

Tabla 2. Forma como se disponen los datos para la simulación

Para el caso del ejercicio de la simulación el día uno, está considerado en la fecha 19 de marzo de 2020, día en que empezó la cuarentena en Colombia. RkTeaching es un paquete R que proporciona un complemento para la interfaz gráfica de usuario RKWard agregando menús y diálogos diseñados. Este paquete ha sido desarrollado por Alfredo Sánchez Alberca asalber@ceu.es en el Departamento de Matemáticas Aplicadas y Estadística del CEU San Pablo de Madrid. Un archivo de instalación del paquete, que instalara a RKward, se puede obtener en la página https://aprendeconalf.es/rkteaching/ . El inicio de la aplicación implica darle un nombre al archivo que se va trabajar, la primera visualización es una hoja como la excel (Figura 1)

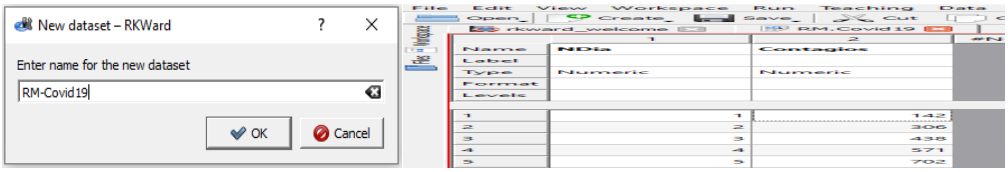

Figura 1. Pantallazo de inicio e ingreso de datos de Rkward.

Todos los procedimientos para realizar cualquier operación en este software, se corren sobre la opción "Teaching", presente en la barra del menú principal. (Figura 2)

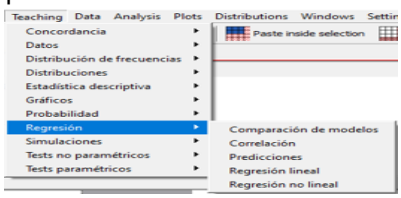

Figura 2. Menú Teaching: Acceso a la opción de regresión

2. RESULTADOS

Para realizar el análisis de correlación múltiple, se realizarán los siguientes procedimientos.

a. Se analiza la matriz de correlación. Si la correlación entre las variables es superior a 0,8 se asegura la existencia de un modelo, al cual se ajustan los datos.
b. Se analizan todos los modelos mediante la comparación de modelos.
c. Una vez se ha elegido el modelo ideal se realiza la predición

a. Matriz de Correlación: A la véntana de la Figura 3, se llega haciendo clik en la opción: Teaching → Regresión → Correlación

Figura 3. Ventana de la matriz de correlación

La opción submit permite encontrar el resultado, que se evidencia en la Tabla 3.

Coeficientes	Contagios	NDia
Contagios	1	0.9864
NDia	0.9864	1

Tabla 3. Matriz de correlación calculada en R-ward

El resultado de la correlación entre las variables, Contagios y NDia, que toma un valor de 0,9864, permiten confirmar la existencia de un buen modelo de correlación, aunque no se muestre con claridad, de cual modelo se trata.

b. Modelos de comparación: para ingresar a esta opción de modelos de correlación, se sigue la ruta:

Teaching → Regresión → Comparación de modelos

Esta ruta permite llegar a la ventana de comparacion de modelos de regresion que se observa en la Figura 4.

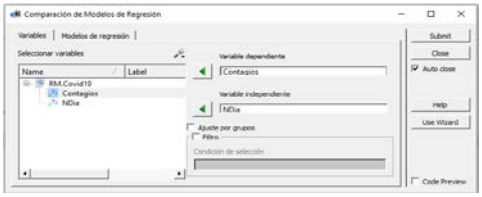

Figura 4. Ventana de comparación de modelos en R-kward

La opción submit, permite observar los diferentes modelos de regresión, lineal y no lineal, con sus respectivos P-valores y R^2. Estos valores representados en la Tabla 4 permiten elegir un modelo ideal que se ajustan a los datos.

Modelo	R^2	P-valor
Cubic	0.9991	< 2.22e-16
Quadratic	0.9953	< 2.22e-16
Potential	0.9923	< 2.22e-16
Lineal	0.973	< 2.22e-16
Exponential	0.8856	< 2.22e-16
Logarithmic	0.7345	2.2405e-13
Inverse	0.2837	0.0002362

Tabla 4. Comparación de modelos en R-Kward

La observación de la Tabla 4, que los modelos cúbicos y cuadráticos son los mejores modelos, de acuerdo a los valores determinados por índice de correlación R^2, el P-valor confirma la hipótesis alternativa de que existe una relación significativa entre las variables (NDia, Contagios). Aunque es mejor la correlación con el modelo cubico, para el ejercicio de la predicción, se escoge el modelo cuadrático.

C. La predicción: La predicción como una profecía, es pensar lo que pasara en el futuro[5] en este caso los modelos, pueden profetizar sobre el número de contagios en el futuro de Colombia. Para realizar el proceso de predicción en el software R-Kward, es necesario, guardar el modelo, una vez se tenga certeza de cuál es el mejor modelo. En este caso se ha elegido tomar el modelo cuadrático. (ver Figura 5)

Figura 5. La predicción en R-Kward

La ejecución de la operación en la opción submit, en el software arroja los siguientes resultados.

Ecuación del modelo

Contagios = 101.388 + 21.4799 Ndia + 2.913 Ndia^2

La ecuación determinada por el software R- Kward evidencia un modelo cuadrático.

Coeficientes del modelo

Coeficiente	Estimación	Error estándar	Estadístico t	p-valor
(Intercept)	101.388	18.82634	5.385433	4.062591e-05

Ndia	21.47993	3.941727	5.44937	3.547687e-05
I(Ndia^2)	2.913014	0.1740089	16.7406	2.03E-12

Tabla 5. Anova del Modelos

La Tabla 5, propone los elementos constitutivos del modelo de regresión cuadrática, su intercepto y lo respectivos coeficientes de las variables, así como sus respectivas significancias que proponen un buen p-valor de correlación lo que verifica la validez del modelo de correlación

Ajuste del modelo

R^2	R^2 ajustado	Estadístico F	p-valor
0.9979065	0.9976739	4290.063	7.723567e-25

Tabla 6. Significancia del modelo

Finalmente, en la Tabla 6, se confirma la hipótesis de que, si existe una correlación entre las variables, lo que supone la validez del modelo y la fortaleza de la predicción

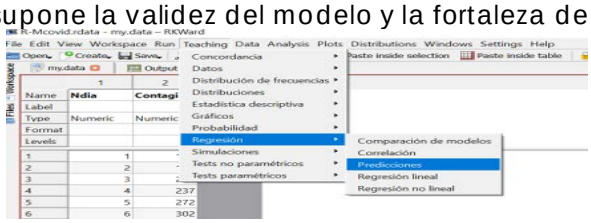

Figura 6. Ejecución de la predicción en R-Kward

Con un modelo significativo, se procede a la predicción, la Tabla 7, muestra algunos resultados predictivos, arrojados por el software. Se busco determinar algunos valores para el día 50, 60, 70, 100, 120, 150, 185, que apropian fechas actuales. La información la refleja la Tabla (FIgura 6)

Ndia	Predicción
50	8457
60	111877
70	158787
100	313795
120	446263
150	668866
185	703773

Tabla 7. Valores de predicción

3. CONCLUSIONES

Consideramos que esta aplicación de R, es útil en los contextos académicos y se acomoda con facilidad a las actividades de los investigadores, el ejemplo, de la utilidad de este software, determina como con cierta comodidad se pueden analizar datos relacionado con los modelos de regresión lineal y no lineal.

La certeza de los resultados propone, la eficacia de la aplicación en este caso en tiempo real en Colombia registra al día de hoy 750.000 casos de contagio, el día 185 con el que se hace la predicción corresponde a esta fecha y propone que existirán 703773 casos activos, lo que no se aleja de esta realidad.

AGRADECIMIENTOS

Agradecer al ministerio de salud colombiano por la disposición pública de los datos en la página y a Sánchez Alberca por la creación de R-Kward, una herramienta que me ha brindado muchas posibilidades para la enseñanza de la estadística.

Referencias

[1] Kim KH. (2020). COVID-19. *Int Neurourol J*. 24 (1): 1-1.

[2] Santabárbara J., López, R. (2019). Actitudes hacia la estadística en residentes de medicina que cursan un posgrado de investigación. FEM: *Revista de la Fundación Educación Médica*, 22(2): 79-83.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

FUNCTIONAL REGRESSION MODELS FOR THE PREDICTION OF CoViD–19

Manuel Oviedo de la Fuente[1] and Manuel Febrero Bande[2]

[1]CiTIUS, University of Santiago de Compostela
[2]Dpt. of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela

ABSTRACT

The motivation for using a functional regression model to predict CoViD19 cases arises from the classic SIR epidemiological model, which in the part specifically dealing with infections proposes the equation: $\frac{dI}{dt} = \beta SI - \gamma I$ where β is the infection rate, γ is the recovery rate, S is the population susceptible to infection and I is the number of infections. Rewriting the above equation we obtain $GR = \frac{dI/dt}{I} = \beta S - \gamma$ which signals that the growth rate of infections (GR) is a function of S. If we extend this idea and discretize the above equation, we can pose the following functional regression model: $GR_{t+h} = f(S^t_{t-l}, GR^t_{t-l}, I^t_{t-l}, \ldots) + \epsilon_{t+1}$ where the notation X^t_{t-l} refers to process X in the interval $[t-l, t]$ and $GR_{t+h} = \frac{I_{t+s} - I_t}{I_t + c}$ with h as the prediction horizon. The function f is the way to link the scalar response with the functional covariates for which the literature on functional data provides several possibilities: i) Lineal Model (FLM)–[1], ii) Spectral Additive Model (FSAM)–[4] and iii) Additive Kernel Model (FKAM)–[2]. All these models are implemented in the fda.usc –[3]- library with which Shiny App is built `http://modestya.securized.net/covid19prediction/`.

The construction of training databases uses information accessible on the web at different resolutions: Data from countries in the repository of the Johns Hopkins University, public data (by region) in Italy, data by Autonomous community in Spain or data by municipality in Catalonia and Madrid. Of course, the availability of quality updated data conditions the selection of the training sample as well as the resolution at which the prediction/estimation/forecast may be made.

Keywords: Functional regression, Predictive modeling, Growth rate, CoViD-19, Shiny application.

ACKNOWLEDGMENT

This work was supported by Project MTM2016–76969–P from the AEI co-funded by the European Regional Development Fund (ERDF), the Competitive Reference Groups 2017–2020 (ED431C 2017/38) from the Xunta de Galicia through the ERDF.

Referencias

[1] Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

[2] Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *TEST*, 22(2):278–292.

[3] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda.usc. *Journal of Statistical Software*, 51(4):1–28.

[4] Müller, H. and Yao, F. (2008). Functional additive models. Journal of the American Statistical Association, 103(484):1534–1544.

VII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 15 de outubro do 2020

ANÁLISIS BIBLIOMÉTRICO BÁSICO CON BIBLIOMETRIX: EL CASO DE LA LONGITUD DEL TELÓMERO EN NIÑOS

Daniel Prieto-Botella[1], Desirée Valera-Gran[1,2], Paula Fernández-Pires[1], Paula Peral-Gómez[1,2], Miriam Hurtado-Pomares[1,2], Alicia Sánchez-Pérez[1,2], Iris Juarez-Leal[1,2], Cristina Espinosa-Sempre[1,2], Eva María Navarrete-Muñoz[1,2]

[1] Department of Surgery and Pathology, Miguel Hernández University, 03550 Alicante, Spain.

[2] Grupo de Investigación en Terapia Ocupacional (InTeO), Miguel Hernández University, 03550 Alicante, Spain.

RESUMEN

El análisis bibliométrico es un método útil para evaluar tendencias y vacíos de conocimiento sobre un tema en la literatura científica. En ocasiones este análisis puede ser un tanto tedioso, especialmente en la recopilación y extracción de la información bibliográfica para el posterior análisis. La herramienta Bibliometrix de R fue diseñada especialmente para realizar análisis bibilométricos partiendo de la información recolectada por las bases Web of Science o Scopus. Para ilustrar la aplicación de esta herramienta y su facilidad de uso, utilizaremos los datos de un estudio bibliométrico sobre la longitud del telómero en niños con bibliografía recuperada de Web of Science. En este trabajo también explicaremos la posibilidad de crear gráficos de redes de palabras clave integrado en esta librería.

Palabras y frases clave: análisis bibliométrico, bibliometrix, bibliometría, R

1. INTRODUCCIÓN

En la actualidad, la producción científica crece exponencialmente cada año, duplicándose el número de artículos publicados cada 10-15 años [1]. El análisis bibliométrico o bibliometría, es una herramienta que utiliza métodos estadísticos para evaluar cuantitativamente la trayectoria documental científica a nivel general o de un área específica, analizando los indicadores de producción e impacto como las citaciones o la colaboración entre países. Este profundo análisis posibilita la detección de tendencias globales y vacíos de conocimiento en la literatura científica [2].

Una de las dificultades para llevar a cabo análisis bibliométricos es la recolección y extracción de la información bibliográfica de las bases de datos para su posterior análisis. Para facilitar este tipo de análisis surge la librería *bibliometrix* del software R, la cual sigue un enfoque de programación funcional, permitiendo automatizar diferentes análisis cuantitativos para obtener información bibliométrica útil como la clasificación de autores, fuentes, países o términos más usados de una forma sencilla. Además, integra funciones novedosas de visualización gráfica de los datos como las redes de colaboración entre países o de concurrencias de palabras clave. Asimismo, esta librería soporta diversos mecanismos de importación para las principales bases de datos (SCOPUS, Web of Science, PubMed, Digital Science Dimensions y Cochrane) [3].

2. USO DE BIBLIOMETRIX EN LA BIBLIOMETRÍA DE LA LONGITUD DEL TELÓMERO EN NIÑOS.

Hace unos meses publicamos un artículo bibliométrico sobre la longitud del telómero en niños (https://pubmed.ncbi.nlm.nih.gov/32604805/) el cual utilizaremos para explicar de forma práctica la potencialidad y facilidad de uso de la librería de R bibliometrix. Para llevar a cabo la bibliometría utilizamos la base de datos de Web of Science combinando las palabras claves "Telomere length" AND "Child". Esta búsqueda nos devolvió 840 artículos que extrajimos en formato BibTex utilizando la herramienta de la propia base. A continuación, se muestra la sintaxis utilizada para cargar la librería bibliometrix y la importación de la información de BibTex mediante la creación del vector "file" a R:

```
library(bibliometrix)
file <- c ("D:/base1.bib", D:/base2.bib")
```

Posteriormente, se procedió a convertir este vector en datos bibliográficos mediante la función *convert2df* creando un objeto, en este caso "M". Esta función contiene dos argumentos, "dbsource" que corresponde a la base de datos (wos= Web of Science) y "format" que identifica el formato del archivo, en este caso BibTex.

```
M <- convert2df(file, dbsource = "wos", format = "bibtex")
```

Una vez convertida la información realizamos un análisis bibliométrico descriptivo a través de la función *biblioAnalysis* y creamos un resumen de los resultados utilizando la función *summary*. Esta función incluye el argumento "k" que permite personalizar la cantidad de resultados por variable que deseemos obtener. Las variables calculadas por la función *biblioAnalysis* incluyen: tipos de documentos, palabras clave, producción científica anual, manuscritos más relevantes, países más productores, citaciones (autor, país y manuscrito), fuentes más relevantes e información sobre autores.

```
results <- biblioAnalysis(M, sep = ";")
S <- summary(object = results, k = 10, pause = FALSE)
```

Un ejemplo de los resultados obtenidos con la función *biblioAnalysis* se encuentra en la Tabla 1. En esta, se muestran los 5 países más productores en longitud del telómero en niños. Bibliometrix nos ofrece datos sobre el número de publicaciones y su porcentaje, además de otra información complementaria muy útil como los artículos publicados por autores pertenecientes a un solo país (SCP) y por pertenecientes a múltiples países (MCP). En nuestro análisis, estos datos nos permitieron identificar la dominancia de EE. UU. en este campo y la escasa proporción de artículos MCP.

Países	**Número de documentos**	**%**	**SCP**	**MCP**	**MCP Ratio**
EE. UU	300	35,7	206	94	0,3
Japón	59	7,0	49	10	0,2
Inglaterra	59	7,0	39	20	0,3
Canada	43	5,1	28	15	0,3
Alemania	43	5,1	18	25	0,6

Tabla 1. Top 5 países más productores en longitud del telómero y niños.

Esta librería de R, además, ofrece más información como el índice H por autores que puede obtenerse con la función *Hinde*x como se muestra en la siguiente sintaxis. Los argumentos importantes de esta función serían "M" (la base de datos creada anteriormente) y "field", la unidad de análisis categorizada en el análisis descriptivo.

```
authors=gsub(",","",names(results$Authors)[1:5])
indices <- Hindex(M, field = "author", elements=authors, sep = ";", years = 50)
indices$H
```

Utilizando la sintaxis descrita anteriormente obtendremos una lista de los 5 autores con mayor índice H (tabla 2). En esta tabla observamos que nos incluye además del índice h, el índice g, el índice m, el número total de citas (TC), el número de artículos (NP) y el año de publicación del primer artículo (YFP). En la bibliometría que realizamos, esta información nos fue imprescindible para detectar que los autores más productivos tienen una experiencia relativamente reciente en este campo, pero acumulan un gran número de citaciones.

Autor	**h-index**	**g-index**	**m-index**	**TC**	**NP**	**YFP**
Savage SA.	13	27	0,9	1120	27	2006
Lin J.	13	26	1,3	1044	26	2011
Lansdorp PM.	16	22	0,7	1791	22	1999
Alter BP.	12	20	0,6	1051	20	2002
Giri N	11	18	0,8	1007	18	2007

Tabla 2. Autores con mayor índice H ordenados por número de publicaciones.

Otra función muy útil que incorpora la librería bibliometrix de R, es la de crear redes para visualizar información. Estas redes son utilizadas para mapear, agrupar y mostrar datos bibliométricos de una forma elegante y novedosa [4]. En nuestro estudio utilizamos la red de concurrencias de palabras clave con el objetivo de evaluar las diferentes líneas de investigación y la red de colaboración entre países. Para realizar estas redes se utiliza la función *biblioNetwork*. Los argumentos importantes son "analysis", donde se indica el tipo de análisis y "network", referida al tipo de red a crear. Para generar la red se utiliza la función *networkPlot*. El siguiente ejemplo corresponde a la creación de una red de concurrencias de palabras clave que se muestra en la Figura 1:

```
NetMatrix <- biblioNetwork(M, analysis = "co-occurrences", network = "keywords", sep
= ";")
net=networkPlot(NetMatrix, normalize="association", weighted=T, n = 30, Title =
"Keyword Co-occurrences", type = "fruchterman", size=T,edgesize = 5,labelsize=0.7)
```

Esta red nos permitió detectar dos líneas de investigación principales, una centrada en el rol del telómero en procesos biológicos y otra con un enfoque epidemiológico sobre las enfermedades ligadas al telómero en la infancia.

3. CONCLUSIÓN

Bibliometrix es una herramienta útil para realizar análisis bibliométricos, simplificando y automatizando análisis cuantitativos, permitiendo un ágil manejo de grandes volúmenes de información. A su vez, incluye herramientas gráficas elegantes que enriquecen el análisis. Sin embargo, la calidad de los resultados depende de la calidad de la información extraída.

Referencias

[1] Bornmann L, Mutz R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66, 2215-22.

[2] Using Bibliometrics: A Guide to Evaluating Research Performance with Citation Data (325133). 12.

[3] Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics 11(4), 959-975.

[4] Waltman L, Van Eck NJ, Noyons ECM. (2010). A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics 4, 629–35.

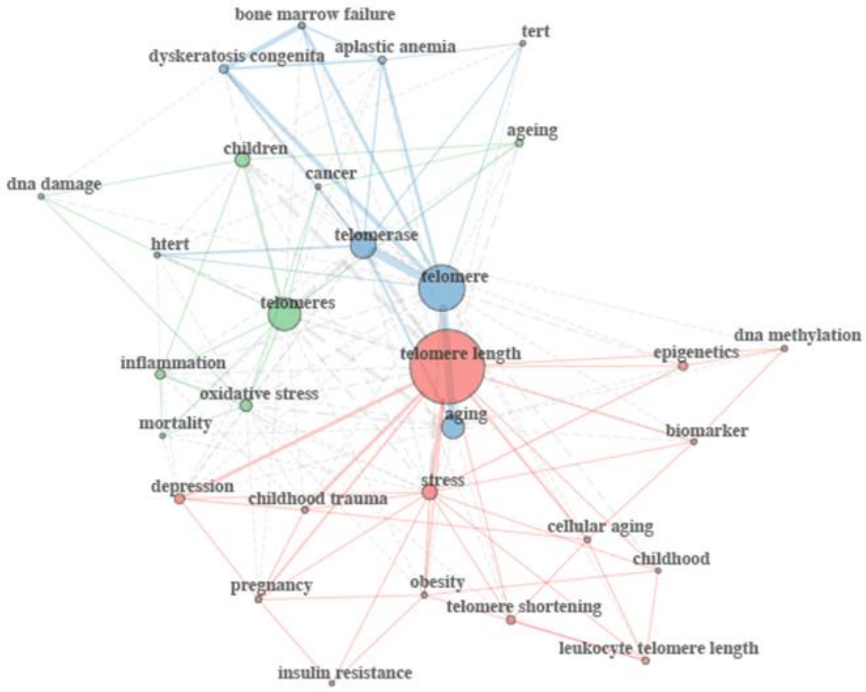

Figura 1. Red de concurrencias de palabras clave en investigación en telómero y niño (n=840 artículos).

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

CÁLCULO DE MAPAS DE RISCO E OCORRENCIA DE INCENDIOS

Marta Rodríguez Barreiro[1], Manuel Novo Pérez[1], Manuel Vaamonde Rivas[1] e María José Ginzo Villamayor[2]

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

RESUMO

Neste traballo preséntase un algoritmo desenvolvido no marco do proxecto Civil UAVs Initiative (CUI), programado en R, *Cálculo de Mapas de Risco e Ocorrencia de Incendios*. Este algoritmo elabora mapas diarios que reflexan o risco de que se produza un incendio nunha área determinada. O índice de risco desenvolvido baséase nun índice existente desenrolado polo ICONA (Instituto de Conservación da Natureza), ó que se lle engade unha modificación para que teña en conta a recorrencia de incendios na zona de interese. O algoritmo devolve un ficheiro en formato *TIFF* que contén un mapa que reflexa o risco de que se produza un incendio na área seleccionada, e que se corresponderá coa hora e a data de execución do algoritmo. Para a creación dos mapas, é necesaria a utilización do paquete **raster**[2].

Palabras e frases chave: Incendios forestais, mapas de risco de incendios, paquete raster.

1. INTRODUCCIÓN

É sabido o gran problema que supoñen os incendios a nivel medioambiental en todo o mundo, sendo Galicia un dos territorios máis afectados de toda Europa. Debido a isto, son numerosos os programas e proxectos que se desenrolan coa finalidade de previr ditos incendios ou de mellorar a forma de loitar contra eles para lograr a súa extinción causando o menor dano posible. Neste contexto, ITMATI traballa coa compañía aeronáutica Babcock no marco do proxecto da Civil UAVs Initative (CUI), desenvolvendo algoritmos orientados a mellorar a xestión das aeronaves de extinción pertencentes á compañía aeronáutica, que ofrece os seus servizos a distintos clientes (entre eles a Consellería de Medio Rural da Xunta de Galicia).
Neste proxecto que se realiza dentro do marco da CUI, ITMATI está a desenvolver varios algoritmos, moitos deles baseados en datos das aeronaves de extinción proporcionados por Babcock. Outros, como o que se presenta a continuación, só precisan unha base de datos con información dos incendios ocorridos nos últimos anos na área de interese, ademais de outra información accesible ó público como o Modelo Dixital do Terreo (MDT) ou o sombreado do mesmo proporcionado polo Sistema de Información sobre Ocupación del Suelo de España (SIOSE). Na actualidade non existe ningunha base de datos pública que conteña información sobre os incendios ocorridos en Galicia, polo que ITMATI utiliza a información recompilada en outro dos algoritmos desenvolvidos (*Algoritmo de detección de incendios*).

2. ESTRUTURA DO ALGORITMO

O algoritmo de *Mapas de Risco e Ocorrencia de Incendios* ten a estrutura que se detalla a continuación:

- En primeiro lugar, obtense información da pendente e a orientación do terreo da área de interese especificada (mediante coordenadas) polo usuario, a partir do ficheiro introducido co MDT. Para isto utilízanse varias funcións do paquete **raster**[2]. Primeiro utilízase a función **raster** para ler o ficheiro *TIFF* (Tagged Image File Format) que contén o MDT. A continuación, úsase a función **terrain** para obter a orientación e a pendente. Por último, utilízase a función **crop** para reducir o terreo á área de interese introducida polo usuario. O seguinte paso é crear unha grella, cuxa extensión coincidirá coa área de interese, e cunha resolución igual a 1km. Para isto, tamén se utiliza a función **aggregate** do paquete **raster**[2] para modificar a resolución do raster e poder crear sobre el un obxecto *SpatialGridDataFrame*. Tamén se le o ficheiro co sombreado do terreo e adáptase á área na que se elaborará o mapa de risco de incendios.

- O seguinte paso é calcular un mapa co índice de risco desenvolvido polo ICONA. Para esto descárganse os datos meteorolóxicos necesarios mediante o servizo *Open Data* de AEMET (Axencia Estatal de Meteoroloxía). Interpólanse estes datos na grella creada previamente e obtéñense dous raster que conteñen a temperatura máxima e a humidade relativa mínima. Utilizando a función **overlay** do paquete **raster**[2] créase, a partir dos dous raster anteriores e unha función que codifica o funcionamento do índice do ICONA, un novo raster que contén o valor da humidade do combustible. Analogamente, a partir dos raster da orientación e da pendente do terreo, úsase a función **overlay** para crear un novo raster coa modificación que hai que aplicar a esta humidade do terreo en función da data e a hora de execución. Por último, coa función **resample** do mencionado paquete modifícanse as resolucións dos raster obtidos para que teñan a mesma, e con **overlay** súmanse para ter o índice de risco creado polo ICONA (pódese ver a construción deste índice en [1]).

- Para continuar, o algoritmo calcula un índice de risco que se basea na ocorrencia de incendios na área de interese (a construción deste subíndice basease nun desenrolo atopado en [3]). Para isto, a partir dos datos dos incendios ocorridos na zona nos últimos 5 anos, créase un raster, cunha resolución de 10km, no que a cada cela se lle da un valor entre 1 e 17 (a escala de valores do índice do ICONA) en función do número de incendios ocorridos nesa cela. Coa función **resample** adáptase a resolución deste raster á do obtido no punto anterior co índice de risco do ICONA. Tamén se crea un segundo raster, coa mesma resolución de 10km, no que se lle asigna un valor entre 1 e 17 en función da gravidade da superficie queimada nos incendios ocorridos (superficie forestal, superficie arborizada non forestal e superficie non forestal). Como no caso anterior, coa función **resample** modifícase a resolución deste raster. Por último, créase un terceiro raster, neste caso coa resolución establecida no raster do paso previo, na que se lle da a cada cela un valor en función da suma dos valores obtidos nos dous raster anteriores. Este último raster será o que conteña o índice baseado na recorrencia de incendios.

- Para finalizar, mediante unha suma ponderada do raster que contén o índice ICONA e o raster que contén o índice baseado na recorrencia de incendios, créase o índice final de risco de ocorrencia de incendios. Mediante a función **writeRaster** do paquete **raster**[2], créase o ficheiro *TIFF* que contén o mapa con este índice representado.

3. APLICACIÓN A DATOS REAIS

Na Figura 1 amósase un mapa de risco de ocorrencia de incendios en Galicia, obtido nunha execución do 14 de Agosto do 2020, ás 12:00 horas.
A resolución do mapa é de 200m. As zonas azuis indican un risco menor, mentres que as zonas en tons vermellos indican un maior risco de que se produza un incendio. Pódese observar que, nese momento da execución, o mapa ten mais tonalidades azuis e amarelas, indicando que non existe un risco xeral. Porén, pódese observar unha zona con tons laranxas na zona da provincia de Ourense, e unha pequena zona en tons vermellos, na costa, indicando un risco elevado de incendios.

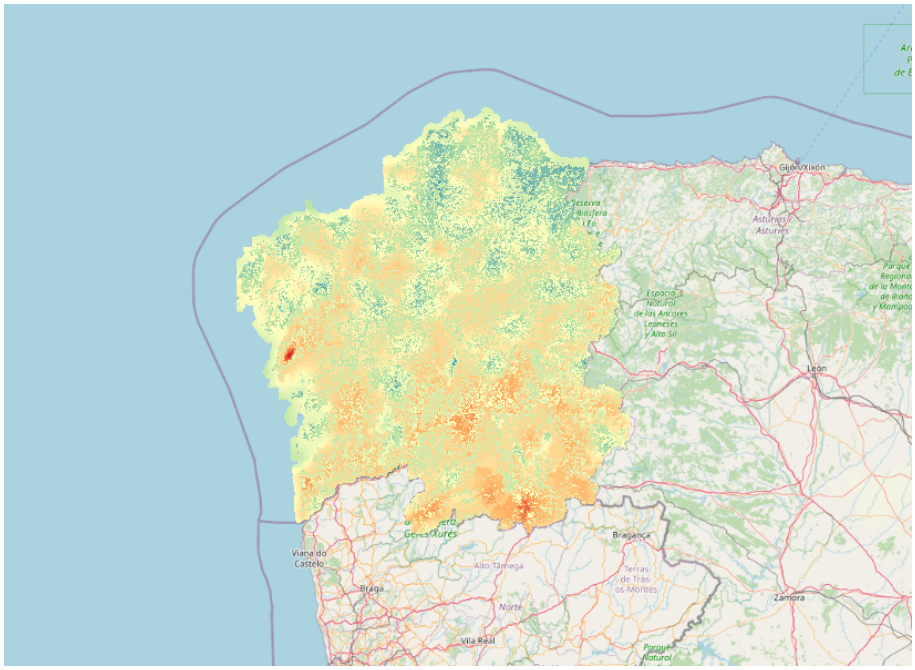

Figura 1: Mapa de risco de incendios en Galicia o 14/08/2020 ás 12:00h

4. CONCLUSIÓNS

O algoritmo desenvolvido proporciona ó usuario un arquivo en formato *TIFF* que contén un mapa que amosa o risco de que se produzan incendios na área de interese de estudio e que é válido para a data e a hora de execución do algoritmo.
O índice desenrolado foi validado mediante técnicas estatísticas, como *Support Vector Machine* (SVM). En xeral, obsérvase que os valores do índice son menores (indicando maior risco) nas zonas que hai incendio que nas que non. Empregando unha mostra de 60 días, entre os anos 2015 e 2019, nos que ocorreu polo menos un incendio, analizando os resultados do SVM e das matrices de confusión pódese concluír que os resultados obtidos son satisfactorios. Concretamente en Galicia, pódese observar que estes resultados melloran sensiblemente que os obtidos ó utilizar unicamente o índice creado polo ICONA.

AGRADECEMENTOS

Os investigadores Marta Rodríguez, Manuel Antonio Novo, Manuel Vaamonde e María José Ginzo agradecen o apoio do proxecto CUI da Axencia Galega de Innovación (GAIN) da Xunta de Galicia e á empresa Babcock International Group plc.

Referencias

[1] Junta de Andalucía. Manual de campo para las operaciones de control y extinción de incendios forestales. Consejería de Medio Ambiente.

[2] Hijmans, R.J. (2019). raster: Geographic Data Analysis and Modeling. R package version 3.0-7. URL https://CRAN.R-project.org/package=raster.

[3] IV Plan General de Defensa contra Incendios Forestales de las Islas Baleares (2015-2024). Publicado en BOIB núm. 56 de 18 de Abril de 2015.

VII Xornada de Usuarios de R en Galicia

Santiago de Compostela, 15 de outubro do 2020

VISUALIZACIÓN DE LA DIRECCIÓN FLUVIAL

Dominic Royé[1]

[1] Departamento de Geografía, Universidade de Santiago de Compostela. Grupo de Epidemiología y Salud Pública

RESUMEN

La visualización de distribuciones de datos circulares, como puede ser la dirección u orientación de variables (p. ej. viento, la costa, el flujo del río, etc.), se suele hacer con histogramas en formato polar. En este artículo se describe usando las direcciones de varios ríos globales, cómo podemos estimar la función de densidad para datos circulares y visualizarla de forma efectiva en R con `ggplot2`.

Palabras clave: visualización, direcciones, ríos, distribución.

1. INTRODUCIÓN

Los ángulos en líneas vectoriales se basan en el ángulo entre dos vértices, y el número de vértices depende de la complejidad, y en consecuencia de la resolución, de los datos vectoriales. Por tanto, puede haber diferencias en usar distintas resoluciones de una línea vectorial, sea de la costa o del río como en este ejemplo. Una línea recta simplemente se construye con dos puntos de longitud y latitud. Relacionado con ello está la fractalidad, una estructura aparentemente irregular pero que se repite a diferentes escalas, de la línea de costa o también del río. La característica más paradójica es que la longitud de una línea costera depende de la escala de medida, cuanto menor es el incremento de medida, la longitud medida se incrementa.

2. PAQUETES

Paquete	Descripción
tidyverse[1]	Conjunto de paquetes (visualización y manipulación de datos): ggplot2, dplyr, purrr,etc.
RQGIS3[2]	Interfaz entre R y QGIS3
sf[3]	Simple Feature: importar, exportar y manipular datos vectoriales
ggtext[4]	Soporte para la representación de texto mejorado con ggplot2
sysfonts	Cargar fuentes en R
showtext	Usar fuentes más fácilmente en gráficos R
circular[5]	Funciones para trabajar con datos circulares
geosphere[6]	Trigonometría esférica para aplicaciones geográficas

Tabla 1: Paquetes requiridos para la visualización y la estimación de la distribución.

3. PREPARACIÓN

Existen dos posibiliades de obtener los ángulos de los vértices. En la primera calculamos el ángulo entre todos los vértices consecutivos. Por ejemplo, imaginémonos dos puntos, Madrid (-3.71, 40.43) y Barcelona (2.14, 41.4).

¿Cuál es el ángulo de su línea recta?

```
bearingRhumb(c(-3.71, 40.43), c(2.14, 41.4))
## [1] 77.62391
```

Vemos que es el de 77º, o sea, dirección noreste. Pero, ¿y si voy de Barcelona a Madrid?

```
bearingRhumb(c(2.14, 41.4), c(-3.71, 40.43))
## [1] 257.6239
```

El angúlo es diferente porque nos movemos desde el noreste al suroeste. Podemos invertir fácilmente el ángulo para obtener el movimiento contrario. La dirección en la que calculamos los ángulos es importante. En el caso de los ríos se espera que sea la dirección de flujo de origen a la desembocadura, ahora bien, un problema puede ser que los vértices, que construyen las líneas, no estén ordenados geográficamente en la tabla de atributos. Otro problema puede ser que los vértices empiecen en la desembocadura lo que daría al angúlo inverso como lo hemos visto antes. Sin embargo, hay una forma más fácil. Podemos aprovechar los atributos de los sistemas de coordenadas proyectados (proyección Robinson, etc) que incluyen el ángulo entre los vértices. Este último enfoque lo vamos usar en este post. Aún así, debemos prestar mucha atención a los resultados según lo dicho anteriormente. En este ejemplo usamos las líneas centrales de los ríos más grandes del mundo, accesible en Zeenatul Basher 2018[7]. Lo primero que hacemos es importar, proyectar y eliminar la tercera dimensión Z.

```
proj_rob <- "+proj=robin +lon_0=0 +x_0=0 +y_0=0 +ellps=WGS84 +datum=WGS84 +un
its=m no_defs"

river_line <- st_read("RiverHRCenterlinesCombo.shp") %>%
                 st_zm() %>%
                    st_transform(proj_rob)
```

En el siguiente paso debemos extraer los ángulos de los vértices. El open software Quantum GIS extrae todos los atributos de los vértices. Es posible acceder a las funciones de QGIS desde R directamente. Para ello, tenemos que tener instalado QGIS en OSGeo4W. El paquete `RQGIS3` nos permite de forma muy fácil usar las funciones del programa en R.

```
# rutas a QGIS
set_env()
# inicio de QGIS Python
open_app()
# buscar herramientas
find_algorithms(search_term = "vertices", name_only = TRUE)
# uso de la herramienta
get_usage(alg = "native:extractvertices")
river_vertices <- run_qgis(alg = "native:extractvertices",
      INPUT = river_line, OUTPUT = file.path(tempdir(), "rivers_world_vertice
s.geojson"),load_output = TRUE)
```

Antes de seguir con la estimación de la distribución de los ángulos, filtramos algunos ríos de interés. Las funciones de la colección `tidyverse` son compatibles con el paquete `sf`.

```
river_vertices <-  filter(river_vertices,
             NAME %in% c("Mississippi", "Colorado",
           "Amazon", "Nile", "Orange", "Ganges", "Yangtze", "Danube",
           "Mackenzie", "Lena", "Murray", "Niger"))
```

Para visualizar la distribución podemos usar, o bien un histograma o un gráfico de densidad. Pero en el caso de estimar la función de densidad de probabilidad debemos hacer uso de funciones especiales para datos circulares. Más detalles estadísticos se explican en el paquete circular. Creamos una función personalizada para estimar la

densidad de probabilidad. Existe una función de optimización para la banda, `bw.nrd.circular()` que se podría emplear aquí.

```
dens_circ <- function(x){
    dens <- density.circular(circular(x$angle, units = "degrees"),
                             bw = 70, kernel = "vonmises",
                     control.circular = list(units = "degrees"))
    df <- data.frame(x = dens$x, y = dens$y/max(dens$y))
  return(df)
}
```

Para finalizar, estimamos la densidad de cada río de nuestra selección.

```
dens_river <- split(river_vertices, river_vertices$NAME) %>%
                    map_df(dens_circ, .id = "river")
```

4. VISUALIZACIÓN

Sólo queda la visualización mediante el famoso paquete `ggplot2`.

```
# descarga de fuente
font_add_google("Montserrat", "Montserrat")
# usar showtext para fuentes
showtext_opts(dpi = 200)
showtext_auto()
```

En el siguiente paso creamos dos objetos con el título y con una nota de pie. En el título estamos usando un código html para dar color a una parte de texto en sustitución de una leyenda. Se puede usar html de forma muy fácil con el paquete `ggtext`.

```
# título con html
title <- "Relative distribution of river <span style='color:#011FFD;'><strong
>flow direction</strong></span> in the world"
caption <- "Based on data from Zeenatul Basher, 20180215"
```

Para la cuadrícula de fondo creamos una tabla con las líneas de fondo del eje x.

```
grid_x <- tibble(x = seq(0, 360 - 22.5, by = 22.5), y = rep(0, 16),
                 xend = seq(0, 360 - 22.5, by = 22.5), yend = rep(Inf, 16))
```

A continuación definimos todos los estilos del gráfico.

```
theme_polar <- theme_minimal() +
      theme(axis.title.y = element_blank(),
            axis.text.y = element_blank(),
            legend.title = element_blank(),
            plot.title = element_textbox(family = "Montserrat",
                         hjust  = 0.5, colour = "white", size = 15),
            plot.caption = element_text(family = "Montserrat",
                                        colour = "white"),
            axis.text.x = element_text(family = "Montserrat",
                                       colour = "white"),
            strip.text = element_text(family = "Montserrat",
                         colour = "white", face = "bold"),
            panel.background = element_rect(fill = "black"),
            plot.background = element_rect(fill = "black"),
            panel.grid = element_blank())
```

Construimos el gráfico final (Figura 1).

```
ggplot() +
  geom_hline(yintercept = c(0, .2, .6, .8, 1), colour = "white") +
  geom_segment(data = grid_x , aes(x = x, y = y, xend = xend, yend = yend), l
```

```
inetype = "dashed", col = "white") +
  geom_area(data = dens_river,
     aes(x = x, y = y, ymin = 0, ymax = y),
     alpha = .7, colour = NA, show.legend = FALSE, fill = "#011FFD") +
  scale_y_continuous(limits = c(-.2, 1), expand = c(0, 0)) +
  scale_x_continuous(limits = c(0, 360),
      breaks = seq(0, 360 - 22.5, by = 22.5), minor_breaks = NULL,
      labels = c("N", "", "NE", "", "E", "", "SE", "",
                 "S", "", "SW", "", "W", "", "NW", "")) +
  coord_polar() +
  facet_wrap(river ~ ., ncol = 4) +
  labs(title = title, caption = caption, x = "") +
  theme_polar
```

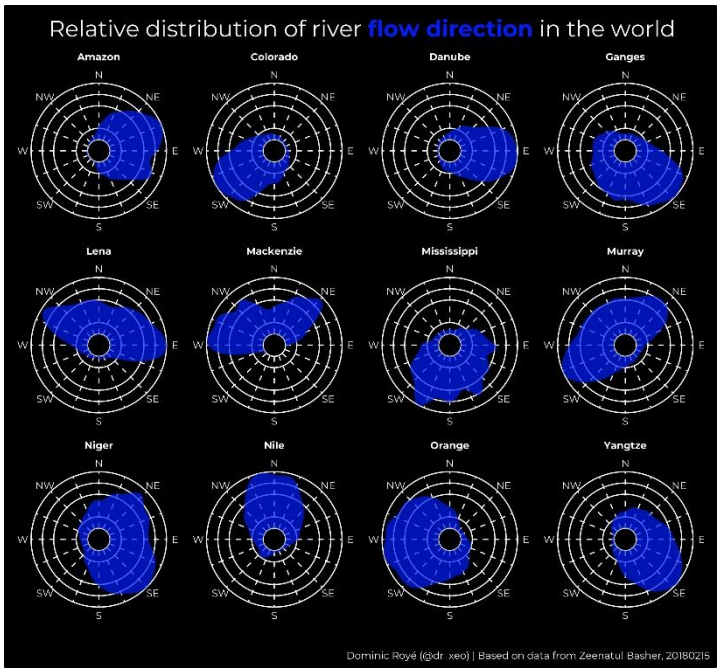

Figura 1: Distribución relativa de la dirección fluvial en el mundo.

Referencias

[1] Wickham H. et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.

[2] Muenchow J. et al. (2017). RQGIS: Integrating R with QGIS for Statistical Geocomputing. *The R Journal*, 9(2), 409-428.

[3] Pebesma E. et al. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1), 439-446.

[4] Wilke C.O. (2020). *ggtext: Improved Text Rendering Support for 'ggplot2'*. R package version 0.1.0. https://CRAN.R-project.org/package=ggtext

[5] Agostinelli C., Lund U. (2017). *R package 'circular': Circular Statistics* (version 0.4-93). https://r-forge.r-project.org/projects/circular/

[6] Hijmans R.J. (2019). *geosphere: Spherical Trigonometry*. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere

[7] Zeenatul B. (2018). *Global High-Resolution River Centerlines*. https://www.sciencebase.gov/catalog/item/5a145fdde4b09fc93dcfd36c

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

R TOOLS TO LINK GAME THEORY AND STATISTICS BY SAMPLING

Alejandro Saavedra-Nieves[1]

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

ABSTRACT

Cooperative game theory deals with the analysis of those conflictive situations where a group of players decides to distribute the profits/costs resulting from their cooperation. In particular, it focuses on defining mathematical tools for proposing allocation vectors that are "acceptable" by the players. A *coalitional value* is a map that associates an allocation vector to every TU-game. The *Shapley value* ([9]) and the *Banzhaf value* ([2]) are probably the most important coalitional values in the literature.

The TU-games with a priori unions incorporate information about the affinities of players. A system of a priori unions is a partition of the player set describing a structure of a priori coalitions. The Shapley value was extended to games with a priori unions (see [5]); this extension is known as the *Owen value*. Besides, [6] proposes the *Banzhaf-Owen value* to extend the Banzhaf value to games with a priori unions.

The calculation of these coalitional values becomes a difficult task when the number of players is large. *Sampling techniques* (see [4]) are an alternative tool for their approximation. In fact, most of coalitional values are averages and then sampling theory ensures good results in their estimation when the set of involved players is sufficiently large. Reference in [3] describes a sampling procedure to estimate the Shapley value, based on simple random sampling with replacement, that is useful in those problems with large player sets. An analogous procedure to approximate the Banzhaf value is introduced in [1]. These procedures are extended in [7] and [8] to those settings with a priori unions. Moreover, the performance of several sampling alternatives is evaluated on a wide collection of examples in literature.

In this talk, we describe the variety of R tools required for dealing with problems as the ones mentioned. It is noteworthy that they can be considered from a statistical point of view and, for this reason, the usage of R software is fundamental. We finally illustrate their performance in real cases as in the estimation of a new system of milk quotas after their suppression and in the estimation of the distribution of the members of the Executive Board of the International Monetary Fund.

Keywords: Games with a priori unions; Coalitional values; Sampling techniques, R.

ACKNOWLEDGEMENTS

Author acknowledges the financial support of *Ministerio de Economía y Competitividad* of the Spanish government under grant MTM2017-87197-C3-2-P and of Xunta de Galicia through the ERDF (*Grupos de Referencia Competitiva*) ED431C 2016-040.

References

[1] Bachrach, Y., Markakis, E., Resnick, E., Procaccia, A. D., Rosenschein, J. S., Saberi, A. (2010). Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, **20**, pp. 105–122.

[2] Banzhaf, J. F. (1964). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, **19**, 317.

[3] Castro, J., Gómez, D., Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, **36**, pp. 1726–1730.

[4] Cochran, W. G. (2007). *Sampling techniques.* John Wiley & Sons 2007.

[5] Owen, G. (1977). Values of games with a priori unions. In Henn, R. and Moeschlin, O. (Eds.): *Mathematical Economics and Game Theory*, 76–88), Berlin: Springer.

[6] Owen, G. (1982). Modification of the Banzhaf-Coleman index for games with a priori unions. In M.J. Holler (Ed.): *Power, Voting and Voting Power*, 232–238. Heidelberg: Physica-Verlag.

[7] Saavedra-Nieves, A., García-Jurado, I., Fiestras-Janeiro, M. G. (2018). Estimation of the Owen value based on sampling. In E. Gil, E. Gil, J. Gil, M. Á. Gil (Eds.), *The mathematics of the uncertain: A tribute to Pedro Gil*, 347–356. Berlin: Springer.

[8] Saavedra-Nieves, A., Fiestras-Janeiro, M. G. (2020). Sampling methods to estimate the Banzhaf–Owen value. To appear in *Annals of Operations Research.*

[9] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, **2**(28), 307–317.

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

NONPARAMETRIC ESTIMATION OF DIRECTIONAL HIGHEST DENSITY REGIONS

Paula Saavedra-Nieves[1] e Rosa María Crujeiras[1]

[1] Universidade de Santiago de Compostela.

ABSTRACT

Density level set estimation theory have been widely considered in literature in the linear setting. However, the specific problem of reconstructing directional density level sets has not been studied in detail. Given a level $t > 0$, it possible to define the directional level set as

$$G_f(t) = \{x \in S^{d-1} : f(x) \geq t\} \tag{1}$$

where f denotes a directional density on the $d-$dimensional unit sphere S^{d-1} for a random vector X. Of course, each $x \in S^{d-1}$ fully characterizes a point in $\theta \in [0, 2\pi)^{d-1}$. Therefore, the definition in (1) can be equivalently established as a subset of points in $[0, 2\pi)^{d-1}$.

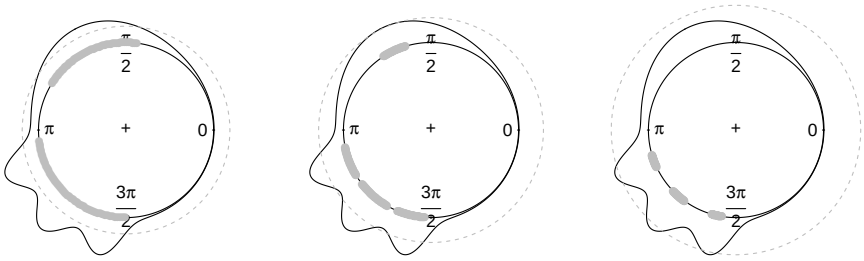

Figure 1: For the quadrimodal circular density represented in black color, $G_f(t_i)$ $(i = 1, 2, 3)$ for t_1 (left), t_2 (center) and t_3 (right) verifying $0 < t_1 < t_2 < t_3$. Equivalently, $L(f_\tau)$ for $\tau = 0.2$ (left) $\tau = 0.5$ (center) and $\tau = 0.8$ (right).

Figure 1 shows the level set $G_f(t)$ in gray color for a circular density and three different values of the level t. The threshold t is represented through a dotted gray line. It should be noted that, if large values of t are considered, $G_f(t)$ will be equal to the greatest modes of the distribution. However, for small values of t, $G_f(t)$ will be virtually equal to the support.

Nonparametric plug-in estimation is the most natural and common way for reconstructing density level sets in the Euclidean setting. Given a random sample $\mathcal{X}_n = \{X_1, \cdots, X_n\} \in S^{d-1}$ of the unknown directional density f, $G_f(t)$ in (1) could be reconstructed as

$$\hat{G}_f(t) = \{x \in S^{d-1} : f_n(x) \geq t\} \tag{2}$$

where f_n denotes, for instance, the kernel density estimator f_n on S^{d-1} proposed in [1]. As in the linear setting, f_n also depends on the selection of the bandwidth parameter which is an essential issue in this estimation approach.

In practice, the specific value of the level t is fully unknown by the user. Therefore, it is necessary to establish an alternative definition of directional level set where

the practitioner chooses the probability content instead of the level. These regions are known as highest density regions in the linear setting. Their generalization for directional data is simple. Given $\tau \in (0,1)$, the $100(1-\tau)\%$ highest density region is defined as the subset

$$L(f_\tau) = \{x \in S^{d-1} : f(x) \geq f_\tau\}$$

where f_τ can be seen as the largest constant such that

$$P(X \in L(f_\tau)) \geq 1 - \tau \tag{3}$$

with respect to the distribution induced by f. Figure 1 shows the highest density region $L(f_\tau)$ in gray color for three different circular densities and three different values of τ. The threshold f_τ is represented through a dotted gray line. Observe that, if large values of τ are considered, $L(f_\tau)$ will be equal to the greatest modes of the density. However, for small values of τ, $L(f_\tau)$ represents the substancial or effective support of the distribution.

Plug-in methods reconstruct the set $L(f_\tau)$ as

$$\hat{L}(\hat{f}_\tau) = \{x \in S^{d-1} : f_n(x) \geq \hat{f}_\tau\}.$$

where $\hat{f}_\tau$ denotes an estimator of the threshold f_τ calculated using numerical integration methods or adapting the approach presented in [2] for the linear case.

In this talk, we will shown how to reconstruct directional highest density regions using the statistical software `R`. Some auxiliary libraries such as `NPCirc`[1] and `Directional`[2] packages, containing useful tools for circular and spherical data, will be also used for our purpose.

Keywords: Directional level set, highest density region, kernel density estimator, bandwidth selectors.

ACKNOWLEDGEMENTS

P. Saavedra-Nieves and R.M. Crujeiras acknowledge the financial support of Ministerio de Economía y Competitividad of the Spanish government under grants MTM2016-76969P and MTM2017-089422-P and ERDF.

References

[1] Bai, Z. D., Rao, C. R. and Zhao, L. C. (1989). Kernel estimators of density function of directional data. *Multivariate Statistics and Probability*, 24–39.

[2] Hyndman, R.J.(1996). Computing and graphing highest density regions. *The American Statistician* 50, 120–126.

[1] `https://CRAN.R-project.org/package=NPCirc`
[2] `https://CRAN.R-project.org/package=Directional`

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

RedBee: SISTEMA DE ANÁLISIS INTELIGENTE DE CIBERATAQUES EN *HONEYPOTS*

Marta Sestelo[1,2], Lilian Adkinson Orellana[1], Borja Pintos Castro[1], Cristian Marques Corrales, Iago López Román[1] y Nora M. Villanueva[1,2]

[1]Gradiant, Centro Tecnológico de Telecomunicaciones de Galicia, Vigo.
[2]Departamento de Estadística e I.O., Grupo SiDOR, Universidad de Vigo.

RESUMEN

RedBee es un sistema que permite analizar y caracterizar de forma automática los ciberataques registrados en *honeypots* Cowrie[1] mediante el uso de técnicas de Machine Learning. Además, permite hacer predicciones y detectar tendencias en los ataques basándose en su evolución temporal.

El sistema dispone de un core analítico, responsable de analizar el contenido capturado por los *honeypots*. Este análisis permite predecir el número de ataques y caracterizar de forma automática toda la información recopilada en forma de *logs*. Para ello, el sistema dispone de los siguientes módulos:

- Módulo de preprocesado: en este bloque se lleva a cabo el proceso de *data wrangling* que se corresponde con la limpieza y procesado inicial de los datos. Aquí se transforman, seleccionan y crean las nuevas variables que alimentarán los modelos de los módulos posteriores. La versión actual del preprocesado puede procesar *logs* con formato de texto no estructurado, así como *logs* en formato JSON. El módulo se encuentra desarrollado en Python.
- Módulo de análisis inteligente: se centra en la caracterización de los ataques. El sistema se basa en algoritmos de aprendizaje no supervisado, en concreto en técnicas *clustering*. Estas técnicas permiten organizar o crear grupos de los ataques registrados en los *honeypots* que presentan características o comportamientos similares. Es importante resaltar que el uso de estas técnicas ofrece un procedimiento automático de agrupación o caracterización de ataques, sin que sea necesario disponer de una base de datos etiquetada y actualizada constantemente. El módulo ha sido desarrollado utilizando la interfaz de Apache Spark [1] para R.
- Módulo predictivo: el foco principal de este módulo es el de predecir el número de ataques diarios que tendrán lugar en el futuro en los *honeypots*. Para ello, se usan Modelos Aditivos Generalizados (GAM) implementados en el paquete de R *mgcv* [2, 3].
- Módulo de visualización: este módulo permite al usuario final visualizar el resultado de análisis descriptivo y el resultado de los modelos aplicados con anterioridad, así como ofrecer resúmenes numéricos y contexto sobre los ataques recibidos. Para su desarrollo se ha utilizado una combinación de Shiny [4] y Kibana[2], una herramienta de visualización de datos integrada con Elasticsearch. Además, los gráficos interactivos de caracterización se han realizado utilizando la librería *plotly* [5].

[1]https://www.cowrie.org
[2]https://www.elastic.co/products/kibana

Palabras y frases clave: *honeypots*, Shiny, ciberataques, modelos predictivos, clustering.

Referencias

[1] Karau, H., Konwinski, A., Wendell, P. Zaharia, M. (2015). *Learning Spark: Lightning-Fast Big Data Analysis*, O'Reilly Media, Inc.

[2] Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

[3] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3–36.

[4] Chang, W., Cheng, J., Allaire, J. J., Xie, Y., McPherson, J. (2019). shiny: Web Application Framework for R. R package version 1.4.0. https://CRAN.R-project.org/package=shiny

[5] Sievert, C. (2018) plotly for R. https://plotly-r.com

VII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 15 de outubro do 2020

MONITORIZACIÓN AUTOMATIZADA DE INCENDIOS

Manuel Novo Pérez[1], Manuel Vaamonde Rivas[1], Marta Rodríguez Barreiro[1] e María José Ginzo Villamayor[2]

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

RESUMO

Neste traballo presentase a integración e interacción ca base de datos e entre eles de dous algoritmos desenvolvidos no marco do proxecto Civil UAVs Initiative (CUI). Os algoritmos, programados en R, execútanse de forma automática para axudar nas labores de extinción de incendios. O primeiro, *Detección de Incendios*, ten unha execución periódica e revisa os datos de voos das aeronaves, que conteñen as descargas de auga realizadas sobre os incendios, para rexistrar na base de datos os incendios activos xunto aos seus perímetros. O segundo algoritmo, *Detección Automática de norias*, é executado de forma automática, cada media hora, para os incendios activos detectados polo algoritmo anterior. O seu obxectivo é estimar os circuitos que realizan as distintas aeronaves entre as zonas de carga de auga e de descarga, denominados norias. A conexión coa base de datos faise mediante conexión HTTP, cos métodos POST e GET, para o que se utiliza o paquete `httr`.

Palabras e frases chave: Incendios forestais, paquete `httr`, BBDD, POST, GET, aeronaves.

1. ALGORITMO DE DETECCIÓN DE INCENDIOS

O obxectivo principal deste algoritmo é proporcionar de forma automática e actualizada os incendios activos. Se ben tamén se pode executar para períodos de tempo concretos, a súa execución en tempo real é a máis interesante. Dispoñer deste rexistro de incendios é de gran utilidade tanto para labores posteriores de análise como para a asistencia na coordinación das tarefas de extinción, pois funciona como base para outros algoritmos. A detección de incendios realízase mediante o acceso a datos de voo das aeronaves, o cal é o único input necesario. A partir das localizacións das descargas de auga efectuadas determínase onde hai un incendio e estímase o perímetro do mesmo. Os principais pasos que segue o algoritmo son:

1. Descarga dos datos de voo mediante o método de HTTP, POST, co paquete `httr` ([3]).
2. Organización e selección dos datos, que veñen en formato JSON. Utilízase o paquete `jsonlite` ([5]).
3. Detección dos incendios a partir das posicións das descargas de auga. Utilízase o paquete `dbscan` ([4]).
4. Axuste dos perímetros por elipses de mínima área co paquete `tlocoh` ([1]).
5. Rexistro en base de datos dos incendios e os seus perímetros. Tamén se inclúen metadatos coma a hora de inicio e fin, as matrículas das aeronaves que participaron na extinción ou o concello no que ten maioritariamente lugar o incendio. Para isto utilízase o paquete `httr` novamente.

6. Carga en base de datos das posicións de descargas asociadas a cada incendio, de forma independente ao punto anterior. De novo, mediante o paquete `httr`.

O algoritmo debe executarse periodicamente para poder actualizar coherentemente os incendios. Así, de cada novo incendio que atopa, xérase un ID de incendio que o identifica e serve para asociarlle a información correspondente, coma por exemplo as descargas. Nas execucións posteriores faise un seguimento sobre cada incendio que permite determinar se está finalizado, se se fusionou con outro próximo ou simplemente actualizar o perímetro e demais información asociada.

2. ALGORITMO DE DETECCIÓN AUTOMÁTICA DE NORIAS

Teoricamente, durante a súa participación nas labores de extinción, as aeronaves realizan o seu traballo seguindo uns circuítos predefinidos (habitualmente, elípticos), denominados norias, entre a zona de carga e a zona de descarga de auga. Na práctica, a definición de norias é bastante complexa, pois estas varían co tempo e as distintas aeronaves poden compartir norias ou non. O obxectivo é resolver o problema de detección e estimación das norias nun incendio, así como das aeronaves que participan en cada unha delas. Ademais, tamén proporciona as zonas de carga e descarga de auga. Este algoritmo depende no anterior, no sentido de que se executará para os incendios activos detectados polo algoritmo de detección de incendios, e utiliza a información de descargas de auga que se garda na base de datos.
O algoritmo de detección automática de norias é largo e complicado, así que neste traballo non se profundará nel en detalle, senón na súa automatización e integración co algoritmo anterior e a base de datos. Os pasos resumidos do algoritmo son:

1. Dado o ID do incendio, obtención das descargas asociadas ao mesmo mediante o método de HTTP, GET, co paquete `httr`.

2. A partir da posición e hora das descargas, determinación da rexión e período temporal no que descargar datos de voo. A descarga realízase de novo mediante o paquete `httr`.

3. Organización e selección dos datos, que veñen en formato JSON. Utilízase o paquete `jsonlite`.

4. Aplícación do algoritmo de detección de zonas de carga, que permite determinar as mesmas a partir das posicións, altitude con respecto ao nivel do mar e velocidade. Utilízase o paquete `dbscan`.

5. Identificación, a partir das zonas de carga, das traxectorias individuais carga-descarga-carga que realiza cada aeronave para cada punto de carga (pode ter varios na mesma execución).

6. Suavización das traxectorias mediante B-Splines periódicos. Utilízase o paquete `pbs` ([6]).

7. Comparación das traxectorias de cada aeronave para cada punto de carga e posterior determinación das norias e asignación traxectorias corresponden a cada unha.

8. Estimación do circuíto que define a noria. Para avaliar á tendencia central das curvas utilízase a profundidade modal con distancia Hausdorff ([2]).

9. Carga en base de datos das norias, xunto ás súas zonas de carga e descarga, e outros metadatos coma as aeronaves que seguen a noria, capacidade potencial ou lonxitude. Cada noria gárdase individualmente, asignada ao ID de incendio. De novo, mediante o paquete `httr`.

3. EXEMPLO DE APLICACIÓN

Amósase un exemplo da aplicación destes algoritmos nun incendio que tivo lugar este verán na provincia de Ourense. En primeiro lugar, na Figura 1 móstrase o perímetro de incendio detectado polo algoritmo de Detección de Incendios sobre un mapa de índice de vexetación de diferencia normalizada (NDVI). As zonas escuras correspóndense con zonas que arderon. Tanto o perímetro coma os puntos asociados a descargas rexístranse na base de datos, todo asociado ao índice identificador do incendio:
Unha vez detectado e rexistrado o incendio na base de datos e prográmase a execución do Algoritmo de Detección Automática de Norias para o identificador deste incendio cada media hora. Na

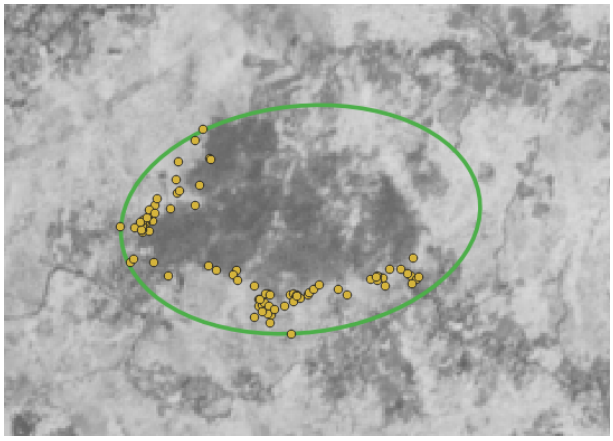

Figura 1: Perímetro do incendio (en cor verde) e as descargas (en amarelo) realizadas nel sobre o NDVI na zona onde ocorreu o incendio.

Figura 2 represéntanse as norias estimadas para este incendio en varias execucións consecutivas. A primeira, ás 17:00, conta cunha única noria correspondente a unha aeronave lixeira. Na execución seguinte, das 17:30, conviven a noria anterior xunto a unha nova correspondente a unha aeronave media. A nova noria ten un punto de carga nunha masa de auga diferente, o cal probablemente se deba ás distintas características de aeronaves lixeiras e medias. Finalmente, ás 18:00, a aeronave media sustitúe completamente á lixeira, que volve á base. As zonas de carga detéctanse correctamente nas masas de auga e as de descarga nos frontes de lapa. En canto ás norias, en todos os casos estímanse traxectorias realistas.

Figura 2: Execucións do algoritmo de detección de norias para as 17:00 (superior esquerda), 17:30 (superior dereita) e 18:00 (inferior). Represéntanse con frechas as norias, con liñas as traxectorias entre cargas, con puntos as posicións de voo, en vermello as zonas de carga e as de descarga en verde. A cor azul correspóndese ca aeronave lixeira e a rosa ca aeronave media.

Tanto as norias como as zonas de carga e descarga son gardadas en cada execución na base de datos, ligadas ao índice do incendio novamente. Durante o traballo no incendio, dende que este se

detecta, continuamente se executan os algoritmos para actualizar toda a información na base de datos e ter unha representación a tempo real tanto do incendio como dos esforzos as aeronaves de extinción.
Para a comunicación ca base de datos utilízase o paquete `httr`, que mediante as funcións `POST()` e `GET()` permite utilizar os métodos HTTP do mesmo nome. O primeiro permite intercambiar información co servidor, xa que se lle adxunta un `body`, en formato JSON, que pode conter tanto información para gardar no servidor coma parámetros de entrada que especifiquen a resposta por parte do servidor. A petición web tamén pode devolver datos en formato JSON, que é a forma en que se obteñen os datos de voo ou demais información gardada na base de datos. A principal diferencia que presenta o método GET é que non permite levar un `body` adxunto. É útil, non obstante, cando só se precisa obter información do servidor e os parámetros de entrada son sinxelos. Neste caso incorpóranse xunto á chamada ao servicio como argumentos.
En ambos casos se precisan credenciais, que se engaden á petición para asegurar una conexión segura.

4. CONCLUSIÓNS

Os algoritmos desenvolvidos permiten facilitar o traballo dos servicios de extinción e proporcionan información difícil de obter. En concreto, o algoritmo de *Detección de Incendios* permite rexistrar os incendios sen a intervención do usuario e facer unha estimación dos perímetros dos incendios a partir das descargas. Unha vez feito o anterior, o algoritmo de *Detección automática de norias* permite, de novo sen intervención do usuario, coñecer as rutas que seguen as aeronaves e resumir o traballo no incendio, que en moitas ocasións son difíciles de determinar con precisión durante unha actuación.

AGRADECEMENTOS

Os investigadores Marta Rodríguez, Manuel Antonio Novo, Manuel Vaamonde e María José Ginzo agradecen o apoio do proxecto CUI da Axencia Galega de Innovación (GAIN) da Xunta de Galicia e á empresa Babcock International Group plc.

Referencias

[1] Lyons, A., Getz, W., and the R Development Core Team (2019). T-LoCoH: Time Local Convex Hull Homerange and Time Use Analysis. R package version 1.40.07.

[2] Cuevas, A., Febrero-Bande, M., Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics 22, 3, 481-496.

[3] Wickham, H. (2019). httr: Tools for Working with URLs and HTTP. R package version 1.4.1. https://CRAN.R-project.org/package=httr

[4] Hahsler M., Piekenbrock M., Doran, D. (2019). "dbscan: Fast Density-Based Clustering with R." Journal of Statistical Software, 91(1), 1-30. doi: 10.18637/jss.v091.i01 URL: https://doi.org/10.18637/jss.v091.i01).

[5] Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

[6] Wang, S.(2013). pbs: Periodic B Splines. R package version 1.1. https://CRAN.R-project.org/package=pbs

AUTORES

VII XORNADA DE USUARIOS DE R EN GALICIA

```
y<-rnorm(12)
x<-1:12
plot(x,y,xaxt="n",cex.axis=0.8,pch=23,bg="gray"
col="black",cex=1.1,main="Uso de 'lines'
para dibujar una serie",cex.main=0.9)
axis(1,at=1:12,lab=month.abb,las=2,cex.axis=0.8
lines(x,y,lwd=1.5)
```

> ORGANIZA

CIXUG

> COLABORA

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

ISBN 9 788409 242733