X XORNADA DE USUARIOS DE R EN GALICIA

18 de outubro de 2023

LIBRO DE RESUMOS

> ORGANIZA

> PATRONCINAN

X XORNADA DE USUARIOS DE EN GALICIA

PROGRAMA E RESUMOS

18 de outubro de 2023

Organiza: Asociación de usuarios de software libre da Terra de Melide

X XORNADA DE USUARIOS DE R EN GALICIA

Editora: María José Ginzo Villamayor

ISBN: 978-84-09-55129-3

Presentación

A Asociación de usuarios de software libre da Terra de Melide (MeLiSA) comprácese en presentar a X 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 e unha mesa redonda para celebrar os Dez anos celebrando a Xornada de Usuarios de R en Galicia. Dos cales oito son convidados e ás outras dez atenderon á chamada de recepción de propostas.
A mesa redonda será moderada por Mª José Ginzo Villamayor (MeLISA e USC), e intervirán:

- Mª Esther López Vizcaíno | Instituto Galego de Estatística
- Salvador Naya Fernández | Universidade da Coruña
- Gael Naveira Barbeito | Consellería de Sanidade
- Miguel Ángel Rodríguez Muíños | Consellería de Sanidade
- Manuel Febrero Bande | Universidade de Santiago de Compostela
- Manuel Oviedo de la Fuente | Universidade da Coruña

Todos eles participaron na primeira I Xornada de Usuarios de R en Galicia, no ano 2013.

Entre os participantes figuran especialistas do Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga), da Xunta de Galicia: diferentes entidades como a Consellaría de Sanidade, Axencia de Modernización Tecnolóxica de Galicia ou o Instituto Galego de Estatística, das tres universidades galegas, de universidades estranxeiras: Universidade Federal Fluminense (Brasil) e da Academia da Força Aérea (Brasil), un profesor de Ensino Medio do IES Pedra da Auga (Ponteareas) e persoal da entidade financeira ABANCA.

Todo isto non sería posible sen o patrocinio de AMTEGA á que agradecemos a súa contribución.

Santiago de Compostela, outubro de 2023
O Comité Organizador

Comité organizador

María José Ginzo Villamayor
Universidade de Santiago de Compostela

Rafael Rodríguez Gayoso
Asociación de usuarios de software libre da Terra de Melide

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

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

X XORNADA DE USUARIOS DE R EN GALICIA

Data

18 de outubro de 2023

Lugar de celebración

Aula Magna. Facultade de Matemáticas (USC)

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 IX Xornada.

Programa

X XORNADA DE USUARIOS DE R EN GALICIA

18 de outubro de 2023

09:10 – 09:30	Sesión de apertura Mª Elena Vázquez Cendón (*Decana da Facultade de Matemáticas*), Salvador Naya Ferández (*Vicerreitor de Política Científica, Investigación e Transferencia – Universidade da Coruña*), María José Ginzo Villamayor (Presidenta de Melisa), Miguel Ángel Rodríguez Muíños (*Saúde Pública – Consellería de Sanidade*)
09:30 – 09:50	Detección de anomalías usando el paquete qcr *Salvador Naya Fernández, Javier Tarrío Saavedra, Miguel Flores e Rubén Fernández Casal	Universidade da Coruña*
09:50 – 10:10	Series de Tempo con R *Manuel Febrero Bande	Universidade de Santiago de Compostela*
10:10 – 10:30	A difusión de estatísticas públicas coas ferramentas que ofrece R *María Martín Vila, Antonio Albo Díaz e Mª Esther López Vizcaíno	Instituto Galego de Estatística*
10:30 – 11:30	Mesa redonda «Dez anos celebrando a Xornada de Usuarios de R en Galicia» *Modera: María José Ginzo Villamayor	Universidade de Santiago de Compostela* *Mª Esther López Vizcaíno	Instituto Galego de Estatística* *Salvador Naya Fernández	Universidade da Coruña* *Gael Naveira Barbeito	Consellería de Sanidade* *Miguel Ángel Rodríguez Muíños	Consellería de Sanidade* *Manuel Febrero Bande	Universidade de Santiago de Compostela* *Manuel Oviedo de la Fuente	Universidade da Coruña*
11:30 – 12:00	**PAUSA**
12:00 – 12:20	Desarrollo de Modelos Predictivos AutoML en Abanca *José Piñeiro Abal e Juan Manuel Mazaira Gómez	ABANCA*
12:20 – 12:40	R aliado da cartografía forestal galega *Laura Alonso Martínez, Andrés Rodríguez Dorna, Julia Armesto e Juan Picos Martín	Universidade de Vigo*
12:40 – 13:00	Usando R para medir a evolución de magnitudes *Alejandro Saavedra Nieves	Universidade de Santiago de Compostela* *Paula Saavedra Nieves	Universidade de Santiago de Compostela, CITMaga*
13:00 – 13:20	R y lme4: Uniendo fuerzas para la Estimación en Áreas Pequeñas *Naomi Diz Rosales e María José Lombardía	Universidade da Coruña* *Domingo Morales González	Universidad Miguel Hernández*
13:20 – 13:40	Avaliación da política de agrupamento escolar segundo a vulnerabilidade social e educativa do alumnado en España *José Manuel Amoedo e Bruno Blanco Varela	Universidade de Santiago de Compostela*
13:40 – 14:00	La (R)evolución de los entornos gráficos de usuario (GUI) para R *Miguel Ángel Rodríguez Muíños	Dirección Xeral de Saúde Pública da Consellería de Sanidade (Xunta de Galicia). Teresa Seoane Pillado	Universidade de A Coruña.*
14:00 – 16:20	**PAUSA**
16:20 – 16:40	R na ciencia: análise de datos – ómicos con R *Adrián Casanova Chiclana	Universidade de Santiago de Compostela*
16:40 – 17:00	Captura de relações semânticas com a similaridade do cosseno: um exemplo em R *Afonso Canosa Rodríguez	IES Pedra da Auga, Conselleria de Cultura, Educación e Universidade (Xunta de Galicia)*
17:00 – 17:20	Uso da linguaxe R como "glue language" para procesamento de expedientes administrativos *Marcos Fernández Arias	Axencia de Modernización Tecnolóxica de Galicia, Xunta de Galicia*
17:20 – 17:40	Design e análise de experimentos com R *Ariel Levy, Eduardo Camilo da Silva, Marcus Antonio Cardoso Ramalho e Mariana Marinho da Costa Lima Peixoto	Universidade Federal Fluminense (Brasil)*
17:40 – 18:00	Unha introdución ó paquete meteospain *Marta Rodríguez Barreiro	Universidade da Coruña, CITMaga. Mª José Ginzo Villamayor	Universidade de Santiago de Compostela, CITMAga*
18:00 – 18:10	**PAUSA**
18:10 – 18:30	Unha xeneaoloxía das teses doutorais en economía e empresa en Galicia co paquete Ggenealogy *José Blanco Álvarez	Universidade de Santiago de Compostela*
18:30 – 18:50	Colorindo mandalas com r: explorando cores e gradientes em curvas planas Joao Paulo M. Santos e Luciane Ferreira Alcoforado	Academia da Força Aérea Brasileira
18:50 – 19:10	Sistemas de recomendación a partir de técnicas de clústering basadas en la estimación tipo núcleo de la densidad Lucía López López	Universidade de Santiago de Compostela. Paula Saavedra Nieves	Universidade de Santiago de Compostela e CITMAga
19:10 – 19:30	Programação linear no plano: uma proposta utilizando ggplot2 *Luciane Ferreira Alcoforado	Academia da Força Aérea Brasileira/Divisão de Ensino*
19:30 – 19:35	Clausura María José Ginzo Villamayor. *Universidade de Santiago de Compostela – Comité Científico*

Índice

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

PROGRAMAÇÃO LINEAR NO PLANO: UMA PROPOSTA UTILIZANDO GGPLOT2

Luciane Ferreira Alcoforado[1]
[1]Academia da Força Aérea Brasileira/Divisão de Ensino

RESUMO

O artigo apresenta uma proposta de ensino e aprendizagem de programação linear (PL) usando a ferramenta ggplot2 do software R. O objetivo é mostrar como a produção de gráficos pode auxiliar na compreensão e na resolução de problemas de PL no plano. O artigo explica os conceitos básicos de PL, como as variáveis de decisão, a função objetivo, as restrições e a região factível. Em seguida, o artigo descreve o funcionamento da ferramenta ggplot2, que permite criar gráficos de alta qualidade e personalizados. O artigo também apresenta exemplos de problemas de PL resolvidos graficamente com o auxílio do ggplot2, destacando as vantagens e as limitações desse método. Por fim, o artigo discute as possibilidades e os desafios de usar o ggplot2 como recurso didático para o ensino de PL. O artigo se baseia em referências bibliográficas sobre PL, visualização de dados e o software R.

Palabras e frases chave: Programação Linear, Método Gráfico, Otimização, ggplot2

1. INTRODUCÃO

A programação linear (PL) é uma técnica matemática que permite modelar e resolver problemas de otimização, ou seja, problemas que envolvem a alocação de recursos escassos para atingir um objetivo, sujeito a certas restrições, (Arenales, 2011; Belfiore & Fávero, 2012). A PL é amplamente aplicada em diversas áreas do conhecimento, como engenharia, economia, administração, biologia, entre outras. Por isso, o ensino e a aprendizagem de PL são fundamentais para a formação de profissionais capacitados para lidar com situações complexas e tomar decisões racionais. No entanto, o ensino e a aprendizagem de PL nem sempre são fáceis, pois exigem o domínio de conceitos abstratos e habilidades lógicas.

Nesse contexto, o uso de recursos visuais pode ser uma estratégia pedagógica eficaz para facilitar a compreensão e a resolução de problemas de PL, especialmente no caso de problemas no plano, ou seja, que envolvem apenas duas variáveis de decisão. Um desses recursos é o método da solução gráfica, que consiste em representar graficamente as variáveis de decisão, a função objetivo, as restrições e a região factível de um problema de PL no plano e determinar qual dos pontos extremos da região factível maximiza ou minimiza a função objetivo. Esse método tem diversas vantagens didáticas, mas também algumas limitações, como a dificuldade de construir gráficos precisos e legíveis manualmente, a restrição ao caso bidimensional e a necessidade de verificar todas as soluções possíveis.

Diante disso, este artigo tem por objetivo propor uma forma didática de abordar o entendimento do método da solução gráfica para Problemas de Programação Linear (PPL) usando a ferramenta ggplot2 do software R. O ggplot2 é um pacote que permite criar gráficos de alta qualidade e personalizados, baseado nos princípios da gramática dos gráficos (Alcoforado, 2021; Wickham, 2016). O artigo mostra como usar o ggplot2 para produzir gráficos que ilustram os problemas de PL no plano e facilitam a aplicação do método da solução gráfica. O artigo também apresenta exemplos práticos resolvidos com o auxílio do ggplot2 e discute as possibilidades e os desafios de usar essa ferramenta como recurso didático para o ensino de PL.

2. O PROBLEMA DE PROGRAMAÇÃO LINEAR

Um problema de programação linear (**PPL**) é um tipo de problema de otimização que busca encontrar a melhor solução possível para uma situação que envolve recursos limitados e objetivos definidos. Utiliza-se de modelos matemáticos que expressam as relações de forma linear entre as variáveis envolvidas no problema, como custos, lucros, produção, demanda, etc. Esses modelos são compostos por um conjunto de variáveis de decisão, que são as variáveis que podem ser controladas pelo tomador de decisão, por uma função objetivo, que representa o que se quer maximizar ou minimizar, e por um conjunto de restrições, que representam as limitações impostas pelo problema.

A estrutura matemática de um **PPL** contendo *n* variáveis de decisão, denotadas por x_1, x_2, …, x_n, pode ser descrita da seguinte forma:

•**Otimizar $z = c_1x_1 + \ldots + c_nx_n$:** função objetivo que indica o valor que se quer maximizar ou minimizar, como o lucro, o custo, a receita, etc.
•**Restrições $a_{i1}x_1 + \ldots + a_{in}x_n$ (<=, = ou >=) b_i, com i=1,2,…,m:** são m inequações lineares das variáveis de decisão, que representam as condições ou limites impostos pelo problema.
•**Região factível:** é o conjunto de todos os valores possíveis das variáveis de decisão que satisfazem todas as restrições. A região factível pode ser representada graficamente por um polígono convexo no plano cartesiano quando o problema apresenta duas variáveis de decisão.
•**Solução ótima:** é o valor das variáveis de decisão que maximiza ou minimiza a função objetivo dentro da região factível. A solução ótima pode ser encontrada pelo consagrado método simplex. Particularmente para problemas de duas variáveis pode-se utilizar o método gráfico.

3. O MÉTODO GRÁFICO, DESAFIOS E SOLUÇÕES COM O GGPLOT2

O método gráfico para problemas de PL é um recurso visual que permite representar e resolver graficamente problemas de PL no plano, ou seja, que envolvem apenas duas variáveis de decisão. Consiste nos seguintes passos:

•Identificar as variáveis de decisão, a função objetivo e as restrições do problema.
•Definir uma escala adequada para os eixos x e y.
•Traçar as retas correspondentes às restrições, usando os coeficientes das variáveis e os termos independentes das inequações.
•Identificar a região factível, que é o conjunto de pontos que satisfazem todas as restrições. A região factível pode ser um polígono convexo, um semiplano, um ponto ou um conjunto vazio.
•Traçar o vetor gradiente e as curvas de nível que são retas perpendiculares ao vetor gradiente.
•Mover as curvas de nível na direção que maximiza ou minimiza a função objetivo, até que ela toque o último ponto da região factível. Esse ponto é a solução ótima do problema.
•Calcular o valor das variáveis de decisão e da função objetivo na solução ótima.

O ggplot2 pode ser usado como recurso didático para o método gráfico de resolução de problemas de programação linear (PL) no plano, pois permite representar graficamente as variáveis de decisão, a função objetivo, as restrições e a região factível de um problema de PL e determinar a solução ótima. No entanto, o uso do ggplot2 também apresenta alguns desafios, que podem ser agrupados em três categorias: conceituais, técnicos e pedagógicos.

Os desafios conceituais dizem respeito à compreensão dos conceitos matemáticos envolvidos no método gráfico e na gramática dos gráficos. Por exemplo, é preciso entender o que são variáveis de decisão, função objetivo, restrições, região factível, solução ótima, vetor gradiente e curvas de nível, bem como o que são dados, mapeamentos estéticos, camadas geométricas, escalas, coordenadas e temas. Além disso, é preciso saber como relacionar esses conceitos entre si e com o problema de PL.

Os desafios técnicos dizem respeito ao domínio das ferramentas computacionais necessárias para usar o ggplot2. Por exemplo, é preciso saber como instalar e carregar o pacote ggplot2 no R, como criar e manipular objetos de dados no R, como construir e modificar gráficos no ggplot2, como resolver possíveis erros que possam surgir durante o processo e como interpretar e apresentar os resultados obtidos.

Os desafios pedagógicos dizem respeito à escolha e à implementação de estratégias didáticas adequadas para usar o ggplot2 como recurso didático para o método gráfico: definir os objetivos de aprendizagem esperados para os alunos, selecionar os problemas de PL mais adequados para ilustrar o método gráfico com o ggplot2, planejar as atividades, avaliar o desempenho e a compreensão dos alunos em relação ao método gráfico e fornecer feedbacks e orientações para a melhoria da aprendizagem.

4. EXPLORANDO OS RECURSOS DO GGPLOT2

Para resolver um problema de PL no plano, considera-se o número de variáveis de decisão $n=2$ e o número de restrições $m > 1$. Cada restrição do problema contém a igualdade, o que significa que uma reta será representada graficamente para cada restrição. A desigualdade definirá em que lado desta reta encontra-se a região factível relativa a esta restrição, conforme figura 1. Como o problema possui pelo menos duas restrições, novas situações semelhantes irão integrar o gráfico até que todas as restrições sejam devidamente representadas, conforme figura 2.

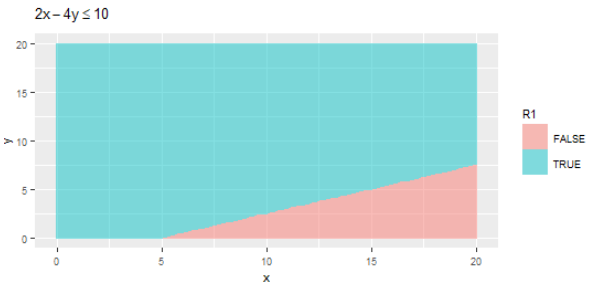


Figura 1: Representação de uma restrição de um PPL.
Fonte: autora(2023).

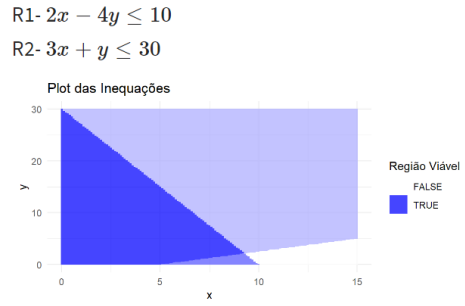


Figura 2: Representação de duas restrições de um PPL.
Fonte: autora(2023).

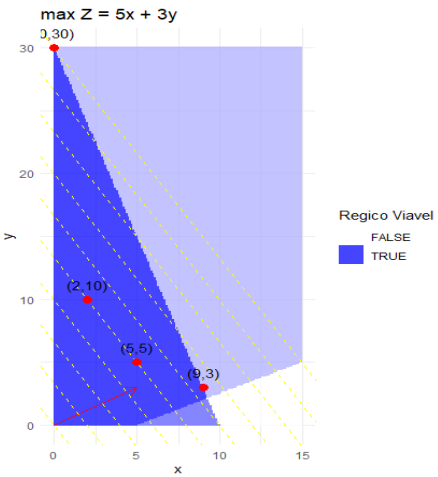


Figura 3: Representação da região viável e curvas de nível.
Fonte: autora(2023).

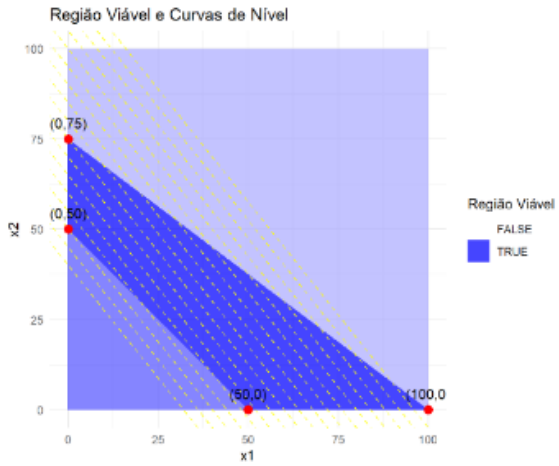


Figura 4: Obter solução que minimize z =10x1 + 8x2.

Para obter a solução ótima ($z^*=90$ com $x_1=0$ e $x_2=30$) é necessário representar as curvas de nível que são obtidas pela função objetivo variando o valor de z, isto é, $z_i = c_1x_1+c_2x_2$, representa a i-ésima curva de nível da função objetivo, linhas pontilhadas da figura 3.

Este recurso permite realizar exercícios com estudantes com o objetivo de treina-los a obter a solução ótima com base nos elementos da representação gráfica, como na figura 4 que se apresenta a região viável e suas curvas de nível, pede-se para obter

a solução ótima para min z = $10x_1 + 8x_2$, cuja resposta deverá ser z*= 400 com x_1=0 e x_2=50.

5. FUNÇÕES NECESSÁRIAS

Inicialmente criamos os objetos necessários para armazenar os dados do problema, os textos em destaque indicam informações que dependem de cada problema:

```
# Data frame com pontos que satisfaçam as inequações
x_vals <- seq(0, lim_x, by = 0.1) # Valores de x
y_vals <- seq(0, lim_y, by = 0.1) # Valores de y
data <- expand.grid(x = x_vals, y = y_vals)
data$ineq1 <- (a11 * data$x + a12 * data$y) <= b1
data$ineq2 <- (a21 * data$x + a22 * data$y) <=
```

1- Função para obter os pontos de intersecção entre os pares de restrições, considerando um objeto *matriz* contendo os parâmetros de cada restrição:

```
intersecao <- c()  # Inicializa o vetor de interseção
# Combinação duas a duas de todas as m linhas da matriz
for (i in 1:(nrow(matriz)-1)) {
   for (j in (i+1):nrow(matriz)) {# Calcula o determinante da matriz formada pelos coeficientes das variáveis x1 e x2
      det <- det(matrix(c(matriz[i,1], matriz[i,2], matriz[j,1], matriz[j,2]), ncol = 2))
      if (det != 0) { # Se o determinante for diferente de zero, calcula os pontos de interseção
         x1 <- det(matrix(c(matriz[i,3], matriz[i,2], matriz[j,3], matriz[j,2]), ncol = 2)) / det
         x2 <- det(matrix(c(matriz[i,1], matriz[i,3], matriz[j,1], matriz[j,3]), ncol = 2)) / det
         intersecao <- c(intersecao, c(x1,x2)) # Adiciona os pontos de interseção ao vetor
      }  }}
```

2- Função para representar a região viável, utilizando o recurso da função geom_tile para produzir uma sobreposição de cores entre as regiões viáveis de cada restrição: `geom_tile(aes(fill = ineq1), alpha = 0.5)`

3- Função para representar o vetor gradiente: `geom_segment(aes(x = 0, y = 0, xend = c1, yend = c2), arrow = arrow(length = unit(0.3, "cm")), color = "red")`

4- Função para obter as curvas de nível: `geom_abline(data = z_data, aes(intercept = intercept, slope = -c1/c2), linetype = "dashed", color = "yellow")`. O intercept é obtido pela função `calc_intercept <- function(z) {  return(z/ c2)}`

6. CONCLUSÃO

Este artigo apresentou uma proposta de ensino e aprendizagem de programação linear (PL) usando a ferramenta ggplot2 do software R. O objetivo foi mostrar como a produção de gráficos pode auxiliar na compreensão e na resolução de problemas de PL no plano, que envolvem apenas duas variáveis de decisão. O artigo explicou os conceitos básicos de PL, como as variáveis de decisão, a função objetivo, as restrições e a região factível. Descreveu como aplicar os recursos do ggplot2 para o desenvolvimento do método de solução gráfica. O artigo também apresentou exemplos de problemas de PL resolvidos graficamente com o auxílio do ggplot2. Por fim, o artigo discutiu as possibilidades e os desafios de usar o ggplot2 como recurso didático para o ensino de PL.

AGRADECIMENTOS

À Academia da Força Aérea pelo apoio ao desenvolvimento do projeto "Aplicativo Web para ensino do método Simplex" (PORTARIA AFA No 87/SPPC) que possibilitou a conclusão deste artigo.

Referencias

[1] ALCOFORADO, L.F. (2021), *Utilizando a Linguagem R: conceitos, manipulação, visualização, modelagem e elaboração de relatórios.* Rio de Janeiro: Alta Books.

[2] ARENALES, M. ET. AL. (2011), *Pesquisa Operacional.* Rio de Janeiro: Elsevier ABEPRO.

[3] BELFIORE, P., FÁVERO, L.P. (2012), *Pesquisa operacional para cursos de administração, contabilidade e economia.* Rio de Janeiro: Elsevier.

[4] WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

X Xornada de Usuarios de R en Galicia

Santiago de Compostela, 18 de outubro do 2023

R, ALIADO DA CARTOGRAFÍA FORESTAL GALEGA

Alonso, L. [1,2], Rodríguez-Dorna, A. [1,2], Armesto, J. [1,2], Picos, J. [1]

[1] Universidade de Vigo, Escola de Enxeñaría Forestal, 36005, Pontevedra, España
[2] CINTECX, Universidade de Vigo, Grupo Xestión Segura e Sostible de Recursos Minerais (XESSMin), 36310, Vigo, España.

RESUMO

A obtención de cartografía forestal e a análise da mesma é imprescindible para o manexo sostible dos bosques. O deseño de metodoloxías empregando R permiten a automatización de procesos e a realización de análises complexas. Neste traballo preséntase a utilización de R para a creación de produtos cartográficos forestais a partir de datos de teledetección a escala rexional (Galicia). Para este proceso foi determinante a ampla comunidade de usuarios de R que presenta un desenvolvemento continuo de ferramentas e solucións que poden ser aplicadas en diversos contextos. A cartografía e os resultados finais de moitos destes análises están sendo incorporados no inventario forestal continuo de Galicia.

Palabras e frases chave: Enxeñaría forestal, manexo sostible dos bosques, R, comunidade de usuarios.

X Xornada de Usuarios de R en Galicia

Santiago de Compostela, 18 de outubro do 2023

Avaliación da política de agrupamento escolar segundo a vulnerabilidade social e educativa do alumnado en España

José Manuel Amoedo[1] e Bruno Blanco-Varela[2]

[1] Grupo de investigación ICEDE, Departamento de Economía Aplicada, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela, Avenida Burgo das Nacións s/n, Santiago de Compostela, España

[2] Grupo de investigación ICEDE, Departamento de Economía Aplicada, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela, Avenida Burgo das Nacións s/n, Santiago de Compostela, España

RESUMO

O coñecemento actual sobre os efectos do agrupamento escolar no rendemento académico é insuficiente e contraditorio. No presente traballo avaliamos o efecto do agrupamento escolar no alumnado segundo a súa vulnerabilidade socioeducativa. Para isto empregamos datos de PISA (2018) para os estudantes en centros españois. A metodoloxía empregada consiste na definición, e cálculo, dun índice sintético de vulnerabilidade socioeducativa e a división do alumnado en base a el e a aplicación de Matching para obter grupos de tratamento e control semellantes. As conclusións indican que esta política ten efectos negativos no rendemento do alumnado e prexudica en maior medida ós alumnos máis vulnerables. Se ben os alumnos con baixa e media vulnerabilidade tamén se ven prexudicados por ela.

Palabras e frases chave: Agrupamento escolar/ vulnerabilidade educativa/ vulnerabilidade social/ política educativa/ PISA/ Matching

1. INTRODUCIÓN

O agrupamento escolar é, na actualidade, unha metodoloxía educativa amplamente estendida no ámbito internacional e español. Sen embargo, o coñecemento sobre os seus efectos no alumnado está lonxe de ser o necesario. De feito, a literatura existente amosa conclusións enfrontadas sobre a súa idoneidade. A vulnerabilidade socioeducativa fai referencia ás barreiras que determinado alumnado debe afrontar para acceder, manterse e beneficiarse do sistema educativo. Concretamente, fai referencia ás características do propio alumno e o seu entorno que conducen a isto.

Dado a natureza do agrupamento escolar o seu impacto semella poder ser asimétrico nos diferentes grupos de estudantes segundo as condicións nas que se desenvolven e, concretamente, da súa vulnerabilidade socioeducativa. O obxectivo deste traballo é o de avaliar o impacto do agrupamento escolar no alumnado con diferentes niveis de vulnerabilidade socioeducativa. Para isto, dividimos o presente traballo en cinco seccións, comezando pola presente introdución. En segundo lugar, definimos o agrupamento escolar, a vulnerabilidade socioeducativa e a relación entre ambos. En terceiro lugar, presentamos a base de datos e a metodoloxía empregada. En cuarto

lugar, presentamos os resultados obtidos tras a aplicación da metodoloxía. Finalmente recollemos as principais conclusións do traballo.

2. AGRUPAMENTO ESCOLAR E VULNERABILIDAD SOCIOEDUCATIVA

O agrupamento escolar é unha metodoloxía docente que consiste na estratificación horizontal do alumnado en función do seu desempeño. Dito agrupamento escolar pode darse dentro das propias aulas, pero tamén dentro de cada centro ou entre centros. A controversia sobre esta metodoloxía de agrupamento escolar refírese ás cuestións de equidade e acceso a unha educación nas mesmas condicións. As posturas máis favorables defenden que estas medidas non supoñen un obstáculo para a mobilidade social. Contrariamente, os detractores do agrupamento argumentan a cuestións relacionadas coa falta de eficiencia e equidade.

A vulnerabilidade socio-educativa é definida como o conxunto de barreiras que conducen a que determinados estudantes presenten problemas ou barreiras para acceder, manterse ou beneficiarse do sistema educativo. O cal tende a condicionar, negativamente, o desempeño do estudante e concluír en problemas como abandono escolar temperán.

A literatura existente sobre o efecto do agrupamento escolar é escasa e contraditoria [1]. Ademais, os traballos existentes non analizan en profundidade se o agrupamento escolar afecta de forma distinta ó alumnado segundo a súa vulnerabilidade socioeducativa. O cal parece ser unha posibilidade dado que a estratificación leva a dividir ó alumnado en grupos moi diferente en canto ós eu grao de vulnerabilidade dada a súa profunda relación co desempeño académico. Polo anterior, parece lóxico pensar que, de feito, o agrupamento de alumnado cun contexto socioeconómico desfavorable pode conducir a un empeoramento do rendemento ou a unha mellora inferior á experimentada polos alumnos con contextos socioeconómicos favorables.

3. DATOS E METODOLOXÍA

Para analizar os efectos do agrupamento escolar no alumnado segundo o seu grao de vulnerabilidade socioeducativa empregamos datos dos estudantes españoles recollidos pola OECD e, máis concretamente, pola edición do 2018 de PISA [2]. A partir dos datos filtrados para alumnos pertencentes a centros educativos españois elaboramos unha base de datos que comprende seis grupos de variables:

1. Variable de tratamento (O centro aplica agrupamento escolar).
2. Variables de rendemento ou resultado (valor medio dos rendementos esperados a nivel global e nas áreas de ciencias, lectura e matemáticas).
3. Variables sobre ás características do estudante (sexo, idade, nacionalidade, repetidor...).
4. Variables sobre o acceso a recursos individuais (espazo propio, computadora, habitación propia, libros...).
5. Variables sobre o entorno socioeconómico do estudante (nivel socioeconómico do fogar, formación dos proxenitores,
6. Variables sobre as características do centro (recursos, tamaños da clase, titularidade do centro, CC.AA...).

O resultado é unha base de datos composta por oitenta variables e 28.215 alumnos. A partir dela medimos o grao de vulnerabilidade de cada alumno seguindo o presentado na sección anterior incluíndo variables dos catro últimos grupos para conformar un índice sintético que contén valor entre 0 e 1. Concretamente, a maior valor do índice maior vulnerabilidade socioeducativa. Na figura 1 presentamos a súa distribución.

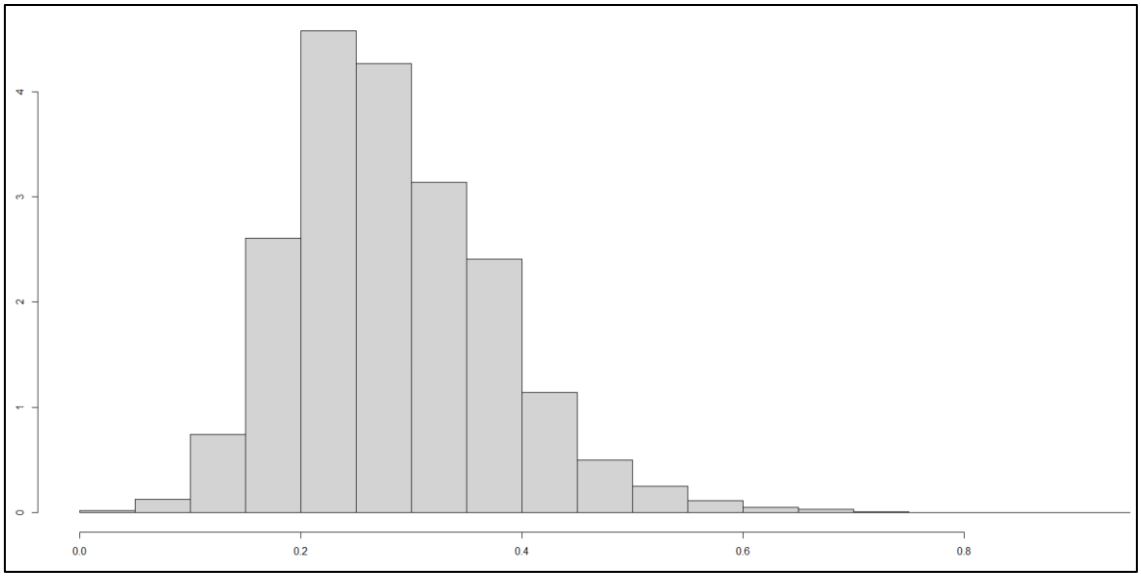


Figura 1: Distribución do índice de vulnerabilidade socioeducativa.

En base á distribución do índice, para dividir ós estudantes en diferentes grupos segundo a súa vulnerabilidade socioeducativa optamos por construír tres grupos. O primeiro está conformado por alumnos con baixa vulnerabilidade (25% con menor valor). O segundo grupo está conformado polos estudantes con vulnerabilidade media (o 50% central). Finalmente, o último grupo confórmase polos estudantes cunha alta vulnerabilidade socioeducativa (25% con maior vulnerabilidade).
A partir dos grupos conformados levamos a cabo o emparellamento mediante Propensity Score Matching coa librería MatchIt desenvolvida e dispoñible para o seu uso co Software libre R [3] empregando as variables dos grupos 3, 4, 5 e 6. Empregamos a metodoloxía coñecida como o Veciño máis próximo cun cociente de 10. A maiores, para algunhas variables clave empregamos distancia exacta co fin de obter un emparellamento idéntico nestes aspectos (Repetidor, Sexo, Inmigrante e CC.AA). O emparellamento levado a cabo amosa boas medidas de balanceo, as cales presentamos na táboa 1.

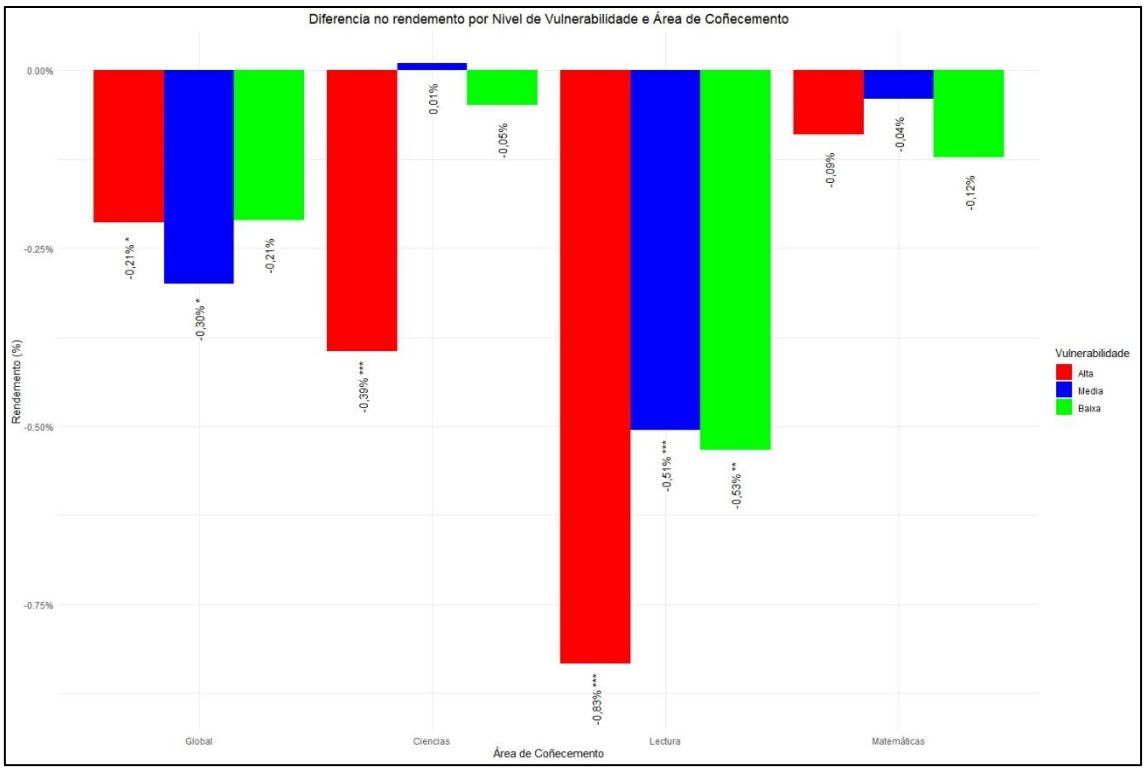
	Alta		Media		Baixa	
	Antes	Despois	Antes	Despois	Antes	Despois
Sesgos medios estandarizados	5,87%	2,54%	5,36%	1,81%	5,59%	1,93%
Pseudo-R^2	0,0432	0,0108	0,0470	0,0074	0,0534	0,0097
Individuos de Tratamento	2.034	2.008	4.036	4.024	2.183	2.149
Individuos de Control	3.915	3.344	7.859	6.432	3.766	3.145

Táboa 1: Medidas de balanceo antes e tras o emparellamento.

A partir dos grupos de tratamento e control obtidos tras o emparellamento calculamos as diferenzas porcentuais entre os primeiros e o segundos grupos co fin de observar o impacto da medida en cada área de coñecemento e en cada tipo de alumnado. A maiores, co fin de comprobar a súa significatividade estatística aplicamos o test-t.

4. RESULTADOS

Os resultados amosan un impacto maioritariamente negativo do agrupamento escolar. O cal, sen embargo, é estatisticamente significativo so en determinados casos. Na Figura 2 presentamos os efectos e a súa significatividade.



Nota: ***p<0,05/ **p<0,10/ *p<0,15
Figura 2: Efectos do agrupamento escolar segundo a vulnerabilidade socioeducativa e a área de coñecemento.

Os resultados amosan como todos os alumnos, independentemente do seu nivel de vulnerabilidade, vense prexudicados polo agrupamento escolar. Tamén o son os alumnos con alta vulnerabilidade en ciencias. Isto refléxase tamén no rendemento global, no cal os alumnos con vulnerabilidade alta e media ven caer o seu rendemento.

5. CONCLUSIÓNS

Neste traballo levamos a cabo unha avaliación dos efectos do agrupamento escolar no rendemento do alumnado segundo a súa vulnerabilidade socioeducativa. Dos resultados obtidos podemos obter as seguintes conclusións. En termos xerais o agrupamento escolar non semella unha boa política, xa que amosa efectos negativos no rendemento dos alumnos. Os alumnos máis vulnerables son os máis prexudicados, polo que podemos concluír que esta política tende a discriminalos. Isto refléxase na súa caída tanto no rendemento global como en ciencias e lectura. Sen embargo, os alumnos con vulnerabilidade media e baixa tamén se ven prexudicados. Concretamente ambos perden rendemento en lectura e, ademais, os primeiros tamén o fan no seu rendemento global.

Referencias

[1] Blanco-Varela, B. (2022). Unha análise socioeconómica da vulnerabilidade en Galicia: o camiño dende a escola até a inserción laboral (Doctoral dissertation, Universidade de Santiago de Compostela). https://minerva.usc.es/xmlui/handle/10347/29389

[2] OECD (2019), PISA 2018 Assessment and Analytical Framework, PISA, OECD Publishing, Paris. https://doi.org/10.1787/b25efab8-en.

[3] Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. Journal of statistical software. https://doi.org/10.18637/jss.v042.i08

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Unha xeneaoloxía das teses doutorais en economía e empresa en Galicia co paquete *ggenealogy*

Jose Blanco-Álvarez[1]

[1]Universidade de Santiago de Compostela, ICEDE Research Group

RESUMO

Empregamos o paquete *ggenealogy* para xerar árbores xenealóxicas amosando a relación entre directores e doutorandos presentando as súas teses no campo da economía e empresa nalgunha das universidades galegas. O traballo amosa os pasos a seguir desde a compilación de datos, pasando polo seu tratamento, interpretación, representación e reflexións sobre a posterior difusión. Alén do uso de R nestes tres últimos pasos, aplícase un enfoque multidisciplinar e integrador empregando outras linguaxes de programación, ben libres como Python, ou privativas como Stata. Aínda que presenta certas limitacións, xustifícase a elección de *ggneaology* en base a súa simplicidade e especificidade para esta tarefa. De cara ó futuro, este traballo presenta posibilidades de ampliación, estudando outras disciplinas científicas, e de mellora, resolvendo as eivas identificadas.

Palabras e frases chave: ggeneaology - Xeneaoloxía - Economía e empresa - Sistema Universitario Galego [SUG] - Teses doutorais - Bibliometría

1. INTRODUCIÓN

A historia da ciencia económica en Galicia estivo marcada por varios fitos históricos. Entre outros, podemos sinalar a implementación da licenciatura en ciencias económicas e empresariais desde o curso 1967/68 e a fundación da Facultade de Economía e Empresa na Universidad de Santiago de Compostela (USC en adiante) poucos anos despois, como o punto de partida da ensinanza *moderna* da economía [3]. A conversión das antigas escolas de comercio en escolas universitarias e a posterior segregación das universidades da Coruña (UDC en adiante) e Vigo (UDV en adiante) nos 90 ampliaron a oferta destes estudos no plano xeográfico e incrementaron o cadro de persoal docente e investigador [1].

En boa medida, estes cadros de persoal nutríronse de profesores formados na propia USC nos anos previos. Se ben nun primeiro momento non foi estrictamente necesario contar cun doutoramento para acceder a prazas de profesorado titular, co paso do tempo estandarizouse un maior nivel de esixencia onde os estudos de doutorado son imprescindibles para obter prazas permanentes na ensinanza universitaria.

Este traballo busca obter información sobre as conexións persoais e profesionais establecidas entre a comunidade científica adicada o campo da economía e empresa en Galicia, mediante a elaboración da árbore xeneaolóxica das teses doutorais presentadas nas tres universidades galegas neste campo. Emprégase un enfoque inclusivo, combinando tres linguaxes de programación: Python para a obtención de datos, Stata para a limpeza e refinado dos mesmos e R para a análise, representación e difusión de resultados. Para este último fin, empregarase fundamentalmente o paquete *ggenealogy* [2] e *Shiny*.

2. OBTENCIÓN DOS DATOS

Existe para o caso de España un repositorio institucional que rexistra a información bibliográfica das teses doutorais defendidas nas universidades do país. Trátase do coñecido como TESEO, dependente do Ministerio de Educación. Entre outros datos, cada entrada neste repositorio inclúe: nome do autor, data de defensa, nome do director e co-directores, membros do tribunal, título da tese, programa de doutorado, área de adscrición e palabras chave (descritores) da tese. Dadas as limitacións dunha interface de consulta francamente mellorable, optouse por recurrir a técnicas de *data scraping* para obter todos os datos dispoñibles[1]. Empregouse para tal fin un programa escrito en Python que, baseándose na librería *Beautiful Soup* accedeu recursivamente a todos os rexistros numerados do 1 ata o 2.200.000 en TESEO[2]. Entre as dificultades a superar estivo a protección desta páxina web contra o acceso masivo, que bloqueaba o acceso desde a nosa IP ó superar os límites fixados. Empregouse un contador que fixaba un intervalo de varios segundos entre cada petición, complementado co acceso a través de *proxies* gratuítos ofrecidos por diversas páxinas web.

Unha alternativa á descarga masiva do TESEO vén dada pola plataforma Dialnet que recolle información das teses doutorais a partir do propio TESEO. Non obstante, a maior protección fronte o acceso masivo e o feito de basearse na mesma fonte a que accedimos directamente, fixéronnos descartar esta opción.

Finalmente, o traballo de Guisán Seijas [1] ofrece unha importante fonte de información para complementar os datos obtidos deste proceso de descarga, xa que existen teses (fundamentalmente aquelas máis antigas) que non aparecen recollidas en TESEO.

3. DATACLEAN

Finalizado o proceso de descarga, obtemos un .csv con un total de 2.200.000 de observacións, correspondentes as teses presentadas entre 1976 e mediados do 2022. Unha parte significativa das mesmas están constituidas por observacións baleiras. Ademais, existe un número importante de entradas con erros e incongruencias, que debemos limpar empregando a linguaxe Stata[3]. Indubidablemente, parte destes erros veñen explicados polo proceso mediante o cal se engaden novas entradas a base de datos. Son os propios doutorandos (con ou sen axuda e supervisión das súas universidades) os encargados de volcar esa información[4].

Finalizado este proceso, seleccionamos aquelas teses defendidas nalgunha das tres universidades galegas, que adscritos a programas de doutoramento ou departamentos de economía e empresa. Para seleccionar aquelas teses que non conteñen información nestes dous campos (as máis antigas), seleccionamos aquelas que teñan a palabra *economía* ou *empresa* entre os seus descriptores.

4. INTERPRETACIÓN E REPRESENTACIÓN DOS DATOS

Do proceso de limpeza de datos obtemos un arquivo con 806 observacións, correspondentes as teses defendidas. Analizando a distribución por anos, obsérvase un incremento sostido das teses en economía e empresa, consecuente co incremento xeral da xeración de coñecemento científico. A USC segue tendo a primacía no número de teses defenidas, con 417 do total, fronte a 210 da UDV e 179 da UDC.

O traballo realizado específicamente con R centráse en representar gráficamente mediante unha árbore xeneaolóxica con *ggenealogy*. Para tal fin, transformárense os datos de tal forma que cada observación inclue o director (*parent*) e o doutorando (*child*). Con este paso obtemos un arquivo con 1038 observacións, resultado de considerar que o fenómeno das codireccións se extendeu na actualidade.

[1] Aínda que este traballo se circunscribe ó caso dun campo específico no SUG, inicialmente obtivéronse os datos de todas as universidades españolas e todas as áreas de coñecemento.

[2] Cada tese está asociada a un id único baseado nunha secuencia de números enteiros comenzando en 1. Non obstante, existen rexistros *baleiros* entre aqueles que conteñen información de teses.

[3] A elección de Python sobre R para o data scraping estaba xustificada por cuestións de practicidade. Neste caso empregar Stata para a limpeza de datos obedece a lóxica da maior familiaridade do autor con este software, pero as mesmas tarefas poderían realizarse de forma igualmente eficiente con R ou co propio Python.

[4] No momento da descarga dos datos, atopamos varias entradas *de proba*. O feito de que algunhas delas xa non estén presentes no momento actual, implica que os responsables de mantemento realizan ás veces comprobacións e depurado da base de datos, o que pode explicar o grande número de ids reservados pero baleiras.

A figura 1 representa un exemplo dunha póla concreta da nosa árbore completa. Neste caso, represéntase a modo de exemplo os doutorandos identificados que tiveron como director de tese a Xose Manuel Beiras Torrado. Sucesivamente, aparecen os doutorandos desta primeira xeración e seguintes.

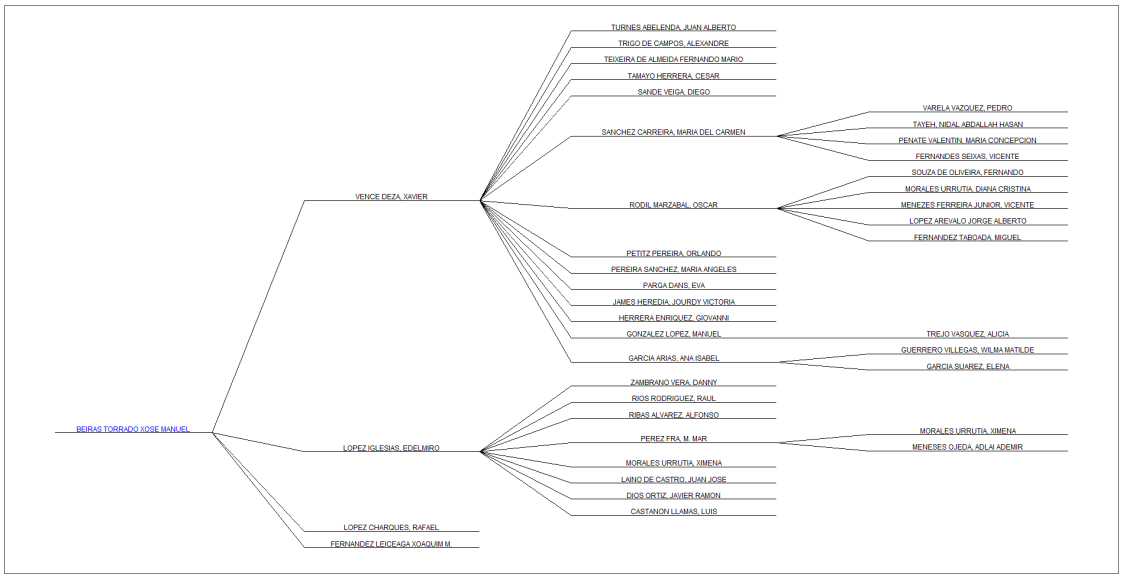


Figura 1: Exemplo de doutorandos identificados con dirección de Xosé Manuel Beiras Torrado

Entre as limitacións máis evidentes deste paquete está o feito de que o gráfico xerado por defecto é moi simple e non inclúe información relativa o ano de defensa da tese (equivalente neste caso a data de nacemento por equivalencia a unha árbore xeneaolóxica normal) ou a afiliación (universidade) do director e doutorando. O auxe das codireccións supón un problema engadido, na medida en que unha mesma persoa pode aparecer en dúas xeracións distintas (ó actuar como director dunha persoa coa que comparte o seu propio director de tese, esta última queda catalogada como filla é *irmá/n* á vez).

5. DIFUSIÓN DOS RESULTADOS

A parte final do proceso consiste en ofrecer os datos resultantes do noso proceso dun xeito accesible ó público xeral. Neste senso, a propia documentación do paquete inclúe referencias a súa integración con *Shiny*. Mediante esta integración, póndense ofrecer ferramentas interactivas de visualización en liña, onde os propios usuarios poden escoller os datos a representar, que aparecen en gráficos e figuras equivalentes ós xerados na execución local. Trátase este dun paso que aínda non foi implementado no noso traballo, pero que sen dúbida ofrece unha gran ferramenta de difusión. Tanto máis cando os datos orixinais incluidos en TESEO ofrecen un acceso difícil e pouco amigable para o usuario.

6. CONCLUSIÓNS

O breve traballo descrito permite amosar as posibilidades de aplicación do paquete e abre novas vías para o estudo da evolución das disciplinas científicas en Galicia. O feito de coñecer as relacións persoais e profesionais establecidas mediante a realización de teses de doutoramento ofrece información acerca das redes e pode ser empregado para reflexionar sobre o proceso de transmisión de ideas.

O enfoque empregado amosa tamén as posbilidades de empregar diversas liunguaxes de programación segundo a conveniencia de cada unha delas para a tarefa concreta, abogando por un enfoque integrador nesta materia.

En canto o paquete *ggenealogy*, comprobouse que a súa especifidade para a tarefa e facilidade de uso son as súas principais vantaxes. Nembargantes, detectaronse varias áreas de mellora que deberán ser abordadas en futuras versións.

AGRADECEMENTOS

Este traballo foi posible en grande medida pola colaboración de Felipe Vieira de Moraes Tavares, que prestou apoio na programación do código de Python para obter os datos do TESEO. Parte da introdución histórica dos estudos de economía e empresa en Galicia baséase nun traballo conxunto (por publicar) cos meus compañeiros Andrés González Rodríguez, Fernando de la Torre Cuevas e Helena Martínez Cabreira. Vaia para todos eles o meu agradecemento.

Referencias

[1] María del Carmen Guisán Seijas. "50 Anos de investigación económica en Galicia, 1967-2017". En: *Revista Galega de Economía* 27.3 (19 de nov. de 2018), págs. 143-166. ISSN: 2255-5951, 1132-2799. DOI: 10.15304/rge.27.3.5627. (Visitado 23-05-2022).

[2] Lindsay Rutter et al. "Ggenealogy: An R Package for Visualizing Genealogical Data". En: *Journal of Statistical Software* 89.13 (29 de mayo de 2019), págs. 1-31. DOI: 10.18637/jss.v089.i13. (Visitado 02-10-2023).

[3] Luis Suárez-Llanos Gómez. "Testemuña duns comezos: actos conmemorativos do corenta aniversario da Facultade de Ciencias Económicas e Empresariais da USC". En: *Revista galega de economía: Publicación Interdisciplinar da Facultade de Ciencias Económicas e Empresariais* 17.2 (2008), págs. 7-16. ISSN: 1132-2799, 2255-5951. URL: https://dialnet.unirioja.es/servlet/articulo?codigo=2767597 (visitado 06-08-2023).

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

CAPTURA DE RELAÇÕES SEMÂNTICAS COM A SIMILARIDADE DO COSSENO: UM EXEMPLO EM R

Afonso Xavier Canosa Rodrigues

[1] IES Pedra da Auga, Ponteareas

RESUMO

A similaridade do cosseno permite capturar relações semânticas entre termos. Nesta comunicação apresentamos um exemplo que mede a similaridade entre topónimos a partir das coocorrências com as classes a que pertencem dentro da tipologia geográfica. Um script em R serve como ferramenta para solucionar todo o procedimento.

Palavras e frases chave: similaridade do cosseno, captura de relações semânticas, hiperonímia, hiponímia, entidades geográficas nomeadas, tipos geográficos

1. INTRODUÇÃO

Segundo o modelo distribucional na semântica, termos relacionados partilham contextos relacionados [1][2]. Temos assim que um modo de capturar relações semânticas é a aplicação de vetores para medir as coocorrências dos termos [3][4]: formamos matrizes de coocorrências e aplicamos uma medida de similaridade (o cosseno) e deste modo obtemos uma medida objetiva da relação.

Para mostrarmos como funciona o modelo, apresento um script em R[1] que aplica a medida do cosseno. Dado que a fórmula pode parecer abstrata, nesta comunicação ponho um exemplo prático para estabelecer a relação entre topónimos (nome próprio) segundo os tipos de entidade geográfica que os classificam (relação semântica de membro de uma classe). Os topónimos estão tirados do texto digitalizado de uma obra clássica, *Peregrinação*, de Fernão Mendes Pinto, na versão da primeira edição de 1614, isto é, com uma língua não normalizada segundo o padrão da atualidade, o qual dificulta aproximações de PLN com as ferramentas convencionais.

2. SELEÇÃO DE ENTIDADES E TERMOS GEOGRÁFICOS E MATRIZ DE COOCORRÊNCIAS

O primeiro passo consiste em criar uma matriz que compute o número de coocorrências das entidades com os termos para os quais queremos estabelecer o grau de relação semântica. Para experimentar com a eficácia da medida,

1 Disponível em:
https://github.com/afonsoxavier/semantics/blob/master/cosine_similarity_results_article.R
Um script mais simples, que soluciona unicamente a similaridade do cosseno sem os procedimentos de tratamento de texto considerados nesta comunicação pode consultar-se em:
https://github.com/afonsoxavier/semantics/blob/master/basic_cosine.R

escolhemos três entidades geográficas e dois tipos geográficos. A tabela 1 amostra o topónimo na forma em que aparece no corpus, a sua frequência, o tipo de entidade geográfica ao que pertence e a sua referência geográfica atual.

EM	Freq.	Tipo geográfico	Referência atual
Çamatra	14	Ilha	Sumatra, Indonésia
Iaoa	26	Ilha	Java, Indonésia
Martauão	35	Cidade (metrópole)	Martabão, Myanmar
Odiaa	19	Cidade (metrópole)	Aiutaia, Tailândia
Pequim	47	Cidade (metrópole)	Pequim, China
Tanixumaa	18	Ilha	Tanegaxima, Japão

Tabela 1: Entidades geográficas mencionadas, tipo (classe) ao que correspondem na atualidade e referência atual

A tabela 2 amostra o número de vezes que cada entidade mencionada coocorre na mesma unidade textual (definida pola presença de um ponto no texto[2]) com os termos *ilha* e *cidade*.

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
CIDADE	1	7	18	17	38	5
ILHA	13	11	1	1	1	11

Tabela 2: coocorrências de entidades geográficas com as expressões CIDADE e ILHA

Uma vez elaborada a matriz, as coocorrências podem ser representadas num sistema de coordenadas (fig. 1) a partir de dois traços, um a abscissa (CIDADE), o outro a ordenada (ILHA).

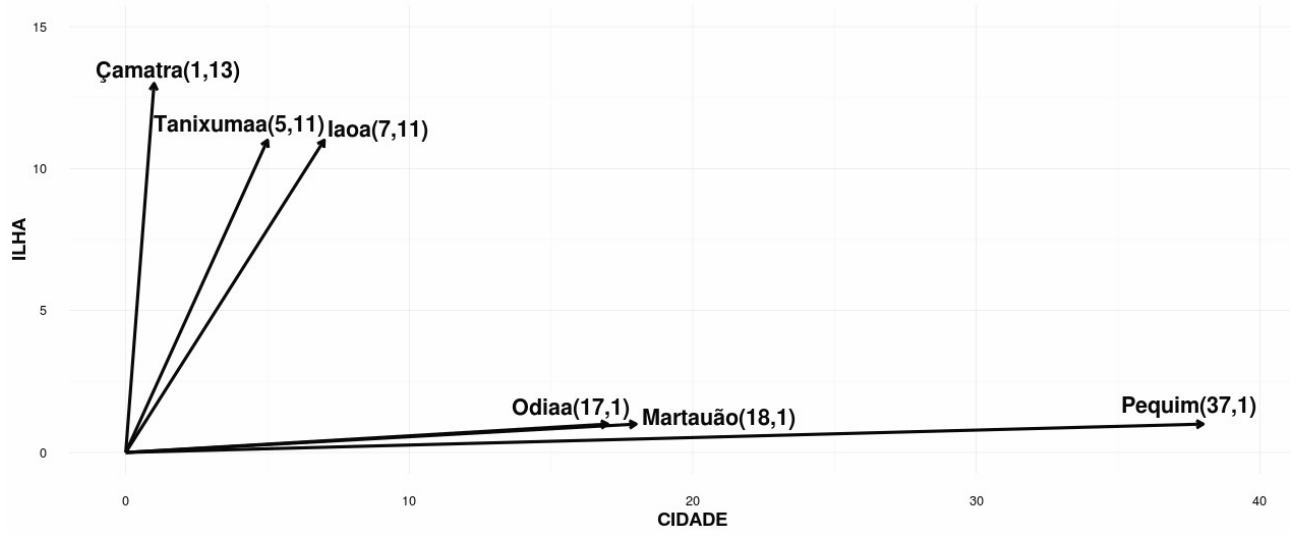


Figura 1: Representação das coocorrências em um diagrama cartesiano.

2 As unidades de segmentação consideradas para as coocorrências processadas neste estudo podem consultar-se no site: https://www.pucau.org. Ex. para *Iaoa*: https://www.pucau.org/?page_id=87&entry=285

Deste modo cada entidade mencionada é definida como um vetor, as suas coordenadas os componentes da posição que podemos representar em forma geométrica com um valor de magnitude e direção.

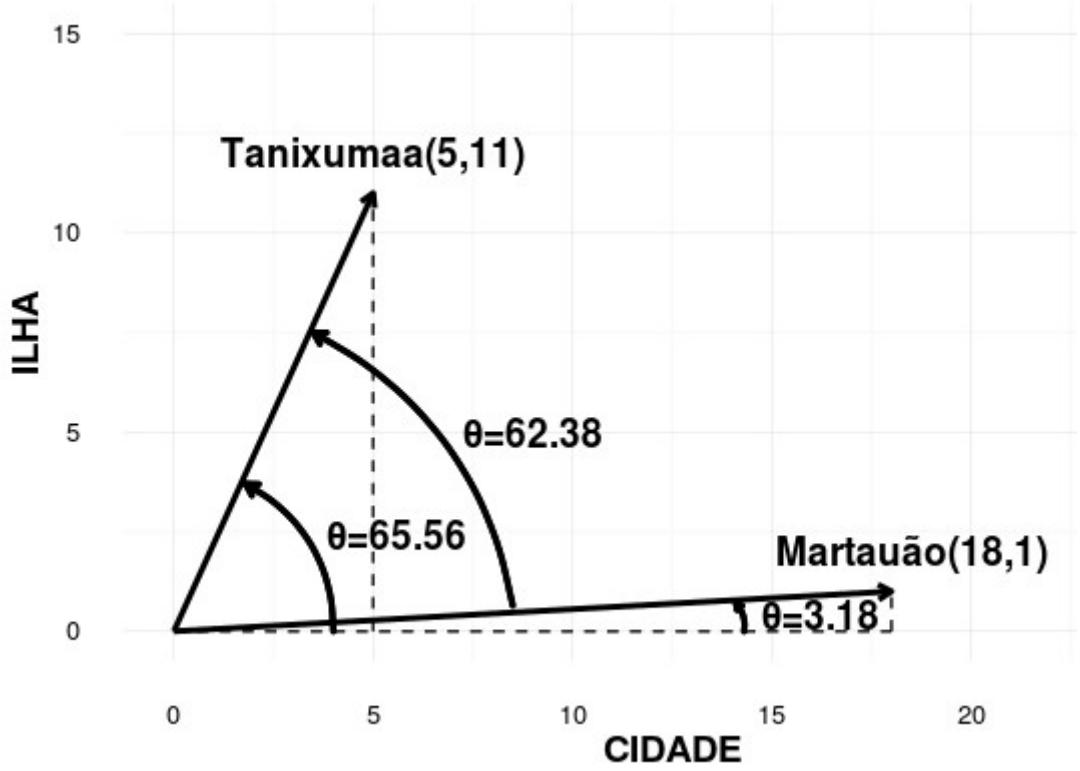


Figura 2: Resolução geométrica dos termos CIDADE E ILHA nas entidades mencionadas *Tanixumaa* e *Martauão*.

3. RESULTADOS

Recolhendo os dados da fig. 2 o script em R (vid. nota 1) opera com as coocorrências para obter os ângulos:

θ (Tanixumaa) = arctan (|11 / 5|) = 65.56°

θ (Martauão) = arctan (|1 / 18|) = 3.18°

Logo a distância semântica para os tipos CIDADE E ILHA das entidades geográficas mencionadas *Tanixumaa* e *Martauão* é:

θ (Tanixumaa) - θ (Martauão)| = |65.56° - 3.18°| = 62.38°

Independentemente da sua magnitude, a direção dos vetores fica entre os 0° e 90°. Obtemos assim uma medida para a similaridade ou distância semântica entre as entidades mencionadas a respeito dos tipos geográficos. Trazemos como exemplo o resultado obtido para a ilha de *Tanixumaa*:

cos(|θ (Tanixumaa) - θ (Çamatra)|) = cos (20.05°) = 0.94

cos(|θ (Tanixumaa) - θ (Iaoa)|)= cos(0.03°) = 0.99

cos(|θ (Tanixumaa) - θ (Martauão)|)= cos(62.38°) = 0.46

cos(|θ (Tanixumaa) - θ (Odiaa)|) = cos(62.19°) = 0.47

cos(|θ (Tanixumaa) - θ (Pequim)|) = cos(64.05°) = 0.44

Os valores mais altos correspondem com a maior similaridade (1 o valor máximo), que achamos com *Çamatra* e *Iaoa*, ambas as duas também ilhas. Quanto mais baixo for o valor do cosseno, menor a similaridade. Para a ilha de Tanixumaa, são os valores obtidos com as cidades *Martauão*, *Odiaa* e *Pequim*.

Um script em R permite-nos calcular facilmente a similaridade entre todos os topónimos considerados (tab. 3). As entidades que pertencem a um mesmo tipo geográfico amostram uma maior similaridade entre elas e uma maior distância com as que não correspondem ao seu tipo geográfico.

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
Çamatra	1.0000000	**0.8823529**	0.1318850	0.1351132	0.1028992	**0.9394222**
Iaoa	**0.8823529**	1.0000000	0.5828468	0.5854906	0.5588836	**0.9902018**
Martauão	0.1318850	0.5828468	1.0000000	**0.9999947**	**0.9995740**	0.4636639
Odiaa	0.1351132	0.5854906	**0.9999947**	1.0000000	**0.9994737**	0.4665474
Pequim	0.1028992	0.5588836	**0.9995740**	**0.9994737**	1.0000000	0.4376085
Tanixumaa	**0.9394222**	**0.9902018**	0.4636639	0.4665474	0.4376085	1.0000000

Tabela 3: Resultados da similaridade do cosseno segundo definida pelos termos ILHA e CIDADE para todas as entidades consideradas no experimento. Em grossa os valores com maior similaridade para cada entidade mencionada.

4. CONCLUSÃO

Mediante um exemplo prático amostramos como, dado um corpus em que as entidades tenham uma ocorrência estatisticamente relevante, a simples coocorrência de termos pode ser utilizada para a captura de relações semânticas. No caso prático estudado aparecem entidades mencionadas em um contexto de linguagem natural, mas com desvios a respeito da norma que dificultam o seu processamento. O problema a resolver consiste em determinar qual é o grau de proximidade entre entidades dados uns tipos geográficos e considerando unicamente as frequências de coocorrência. Nos exemplos estudados, os resultados obtidos amostram a relevância da similaridade do cosseno para medir a afinidade semântica com outras entidades geográficas, oferecendo assim uma solução automática para o relacionamento de topónimos e a sua adscrição a um tipo geográfico.

Referências

[1] Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388-1429.

[2] Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, *9*.

[3] Jurafsky, D., & Martin, J. H. (2015). Vector Semantics. In *Speech and Language Processing* (3rd ed)

[4] Gamallo, P. (2016). Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 1-17.

X Xornada de Usuarios de R en Galicia

Santiago de Compostela, 18 de outubro do 2023

R NA CIENCIA: ANÁLISE DE DATOS -ÓMICOS CON R

Casanova A, Aramburu O, Arana ÁJ, Carreira M, Fernández M, Mejuto N, Pampín M, Quelle-Regaldie A, Sambade IM, Torres D, Valín R, Veiga-Rúa S, Villamayor PR, Blanco A, Cabezas P, Fernández C, Hermida M, Robledo D, Rubiolo J, Bouza C, Pardo BG, Sánchez L, Vera M, Vilas R, Viñas A e Martínez, P.

Departamento de Zooloxía, Xenética e Antropoloxía Física, grupo ACUIGEN, Facultade de Veterinaria, Campus Terra, Universidade de Santiago de Compostela, 27002 Lugo, España.

RESUMO

O persoal investigador do grupo ACUIGEN (GI-1251; USC) traballa na análise bioinformática de diferentes datos -ómicos tanto de especies de interese comercial como desde unha perspectiva de conservación dos recursos naturais. O entorno de R e o seu amplo abano de paquetes permiten as diferentes análises dun xeito máis sinxelo e eficiente. Neste traballo son tratados brevemente tanto as causas do crecente peso de R no campo da bioinformática así como as diferentes ferramentas empregadas no contexto de investigación do grupo.

Palabras e frases chave: Xenómica, transcriptómica, epixenómica, enfermidades de especies comerciais, conservación.

1. INTRODUCIÓN

R é un entorno e unha linguaxe de programación libre e gratuíta deseñada para a realización de análises estatísticas e a produción de gráficos. Desde o ano 1997, trala creación do repositorio CRAN (*The Comprehensive R Archive Network*), as persoas usuarias desta linguaxe foron programando un crecente catálogo de extensións de R, denominadas paquetes, ampliando en gran medida a súa funcionalidade inicial. Na actualidade, o repositorio CRAN conta cuns 20,000 paquetes (R-4.3.1), cando no ano 2013 había ~6,000 (R-3.0). Alén deste repositorio, creáronse outros como *Bioconductor* (https://www.bioconductor.org/), *R-forge* (https://r-forge.r-project.org/) e *GitHub* (https://github.com/) que presentan tamén miles de paquetes de R. Actualmente, con R é posible realizar un abano de análises bioinformáticas cuasi infinito, entre as cales atoparíanse as análises dos datos -ómicos.

2. E QUE SON OS DATOS -ÓMICOS?

Este sufixo é empregado para abranguer a totalidade de entidades biolóxicas, como pode ser o xenoma (i.e., xenómica), transcriptoma, etcétera. (Figura 1). Nos últimos quince anos, o rápido desenvolvemento de tecnoloxías de secuenciación masiva e o seu abaratamento incrementou o volume de datos biolóxicos exponencialmente, supoñendo un desafío para o seu procesamento e posterior interpretación biolóxica.

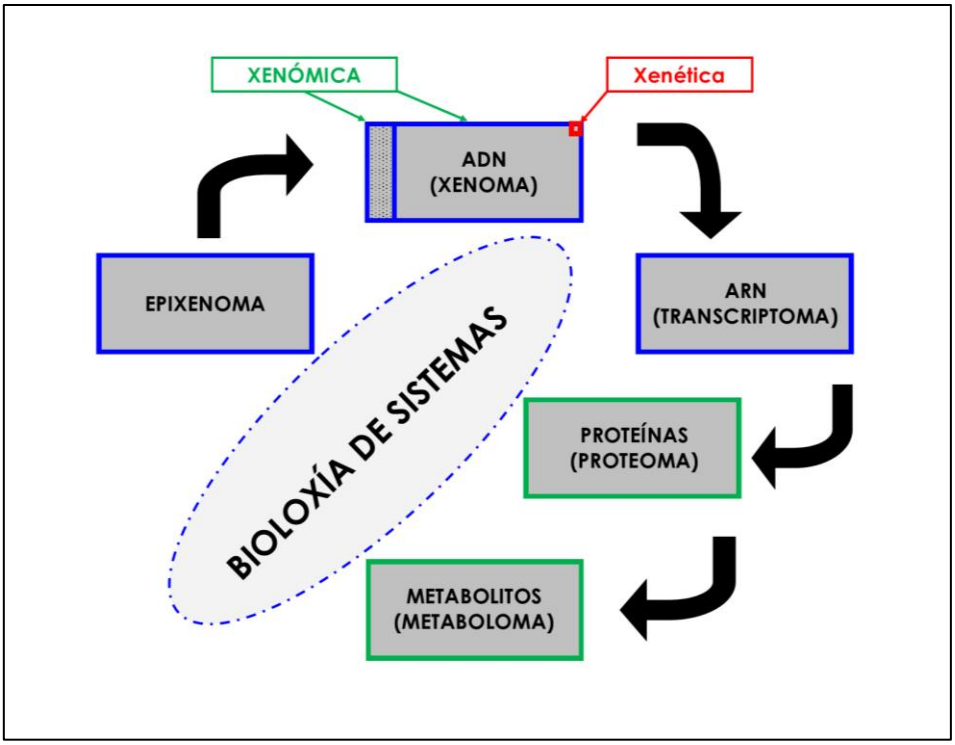


Figura 1: Diagrama das diferentes -ómicas cunha relación biolóxica simplificada. Todas contribúen no campo da bioloxía de sistemas se ben non o completan na súa totalidade. Na parte superior refléxase a diferente escala na que traballan as aproximacións xenéticas e xenómicas, se ben estas últimas non implican traballar coa totalidade da información (ver caixa azul con recheo de puntos). No grupo de investigación ACUIGEN trabállase principalmente na epixenómica, xenómica e transcriptómica, enmarcadas en azul escuro.

Cabe aclarar que traballar a escala -ómica non implica necesariamente traballar coa totalidade da información, por exemplo, traballar cos máis de 2,300,000,000 nucleótidos que ten o xenoma da troita común. Senón cunha mostraxe que poida ser representativa para os obxectivos da investigación. Por exemplo, milleiros ou millóns de marcadores coma Polimorfismos de Nucleótido Único (SNPs) distribuídos ao longo do xenoma en vez duns poucos marcadores altamente informativos como se realiza nas aproximacións xenéticas (recadro vermello, Figura 1). Esta escalada de información biolóxica permite responder con maior seguridade a preguntas previamente expostas, atopar multitude de novas respostas así como explorar vastas rexións do coñecemento que na práctica permanecían inexplorables.

3. POR QUÉ TODOS OS CAMIÑOS LEVAN A R-OMICS?

O emprego de R presenta unha elevada popularidade no análise de datos [1] que se pode ver reflectida no índice TIOBE (https://www.tiobe.com/tiobe-index/) ou no índice de *PopularitY of Programming Language* (PYPL; https://pypl.github.io/PYPL.html). Isto reflíctese nunha forte actividade comunitaria e nun incremento no número de publicacións anuais en diferentes revistas académicas presentando novos paquetes dispoñibles (Figura 2).

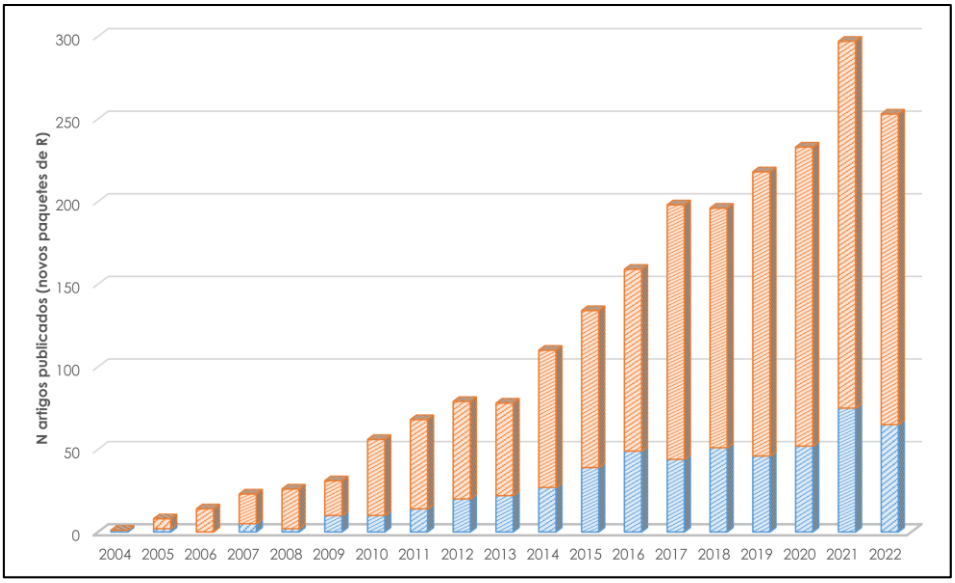


Figura 2: Número de artigos científicos co lanzamento dun novo paquete de R publicados anualmente. En laranxa, artigos publicados en *Open Access*. En azul, artigos publicados con acceso limitado (e.g., subscrición á revista). Base de datos: *Web of Science Core Collection*. Estratexia de procura non estrita en títulos (TI) de publicacións: TI = ("R package").

A súa elevada popularidade en bioinformática tería múltiples factores, dos que destacaríamos catro: (i) R é un programa libre e gratuíto, (ii) funciona practicamente en todos os sistemas operativos, (iii) a transversalización das competencias en programación e (iv) a súa dispoñibilidade nos diferentes centros/recursos de supercomputación tanto públicos coma privados. Nos últimos anos, presentar habilidades de programación é unha competencia valorada no mundo científico *in sensu amplo* (e.g., Número especial da revista Molecular Ecology Resources[1]), así como a inclusión destas habilidades nos plans de estudo de diferentes graos/másters máis aló das carreiras de informática e matemáticas onde se atopaban "restrinxidas" anteriormente[2]. Finalmente, debido ao elevado volume de datos xerados o acceso a recursos de supercomputación para a almacenaxe e análise vólvese fundamental en bastantes proxectos científicos. Neste eido, R atópase instalado en diferentes centros de supercomputación coma o Centro de Supercomputación de Galicia (CESGA), podendo crear cada persoa usuaria a súa libraría persoal de paquetes de R. Deste xeito, aproveitando a serie de métodos para o procesamento paralelo do código que posúe R, pódense empregar ducias ou centos de núcleos de procesamento e elevados volumes de memoria RAM, moi superiores aos dispoñibles a nivel local.

4. ANÁLISE DE DATOS -ÓMICOS CON R

Para xenómica de poboacións existen diferentes paquetes de R para xestionar o abano de formatos no que esta poida estar contida. A modo de exemplos temos os paquetes **radiator** [2] e **GENEPOPEDIT** [3] permitindo os filtrados de marcadores, ordenamento de mostras e a interconversión de formatos con arquivos de datos moi pesados. Para determinar a diversidade xenética, a materia prima da evolución das diferentes poboacións naturais, temos paquetes como **diveRsity** [4] e **genepop** [5]**.** Para avaliar a diferenciación xenética (F_{ST}) entre poboacións podemos usar **genepop** ou **StAMPP** [6]. Para coñecer a estrutura das poboacións podemos empregar **adegenet** [7,8], ou paralelizar o software STRUCTURE [9] empregando **ParallelStructure** [10] usando 100 *cores* no CESGA. Detectar e cuantificar *Runs of Homozygosity* (ROHs) para avaliar a

[1] Special Issue of Molecular Ecology Resources journal (2017): Population Genomics with R. Ligazón: https://onlinelibrary.wiley.com/toc/17550998/2017/17/1

[2] Sirva a modo de exemplo a materia obrigatoria de "Análise xenómica e bioinformática" impartida no Grao en Bioquímica. Ligazón: https://www.usc.gal/es/estudios/grados/ciencias/grado-bioquimica/20222023/analisis-genomico-bioinformatica-17863-17086-2-99236

endogamia das poboacións naturais ou baixo explotación comercial mediante o paquete **detectRUNS** [11], baseado na correlación positiva entre os maiores niveis de endogamia coa lonxitude dos ROHs**.** Clasificar por parentesco os individuos de acuicultura ou gandería usando **related** [12]. Pescudar pegadas de selección no xenoma empregando **pcadapt** [13] e **OutFLANK** [14], así como identificar posibles adaptacións locais mediante o análise entre as frecuencias alélicas a traveso das poboacións e as diferentes variables ambientais empregando **VEGAN** [15]. Implementar estudos de asociacións de xenoma completo (*genome-wide association study*; GWAS), empregando millóns de marcadores xenotipados detectando loci/xenes asociados a algunha característica de interese (e.g., resistencia a enfermidades, crecemento) cos paquetes **GWASTools** [16] e **rMVP** [17]. Con todo, cun enfoque holístico, resulta interesante estudar os efectos de diferentes mecanismos que actúan sobre o ADN (e.g., metilación de histonas), regulando as expresións dos xenes, sen implicar a modificación da secuencia deste (i.e., epixenómica). Nestas metodoloxías empréganse paquetes coma **segmenter** [18] que permite integrar a información provinte de diferentes técnicas: datos de cromatina aberta mediante ATAC-seq + datos de histonas modificadas con marcas de activación/inhibición mediante ChiP-Seq, para determinar o estado no que se atopa o xenoma. Tamén podemos avaliar os datos das variacións epixenómicas dunha especie comercial ante parellas de diferentes condicións experimentais (e.g., presencia/ausencia axente patóxeno) usando **DiffBind** [19,20].

Amais de coñecer a secuencia do xenoma, así como variantes nucleotídicas, epixenómicas, estruturais, etcétera; en moitas investigacións resulta crucial avaliar como se expresan os xenes nas diferentes condicións ambientais e fisiolóxicas, entre os tecidos do organismo estudado. No ámbito da transcriptómica **DESeq2** [21] e **sleuth** [22] serven para cuantificar a expresión diferencial nos diferentes tecidos entre as diferentes condicións de experimentación testadas, determinando que xenes se están a expresar diferencialmente. Traballando con datos cunha resolución superior aos tecidos, podemos utilizar **Seurat** [23], avaliando a expresión cos datos obtidos da secuenciación de ARN unicelular (*Tecnoloxía de single-cell/single nuclei RNA Sequencing*), procesando a información da unidade biolóxica.

Finalmente, non hai que descoidar o aspecto gráfico dos resultados. Podemos visualizar e mapear resultados xenómicos en gráficos atractivos con **RIdeogram** [24], **CMplot** [25] para obter visuais gráficos de Manhattan para resultados de GWAS ou plasmar a expresión diferencial do ARN con **EnhancedVolcano** [26] ou **pheatmap** [27].

5. CONCLUSIÓNS

O uso do entorno de R nas análises -ómicas atópase plenamente consolidado. Seguindo as tendencias actuais, xurdirán novas ferramentas en consonancia cos novos avances, datos e desafíos que xurdan neste dinámico campo.

AGRADECEMENTOS

AC é financiado por un contrato postdoutoral Xunta de Galicia-Campus Terra (2022). IMS é financiada por un contrato predoutoral Xunta de Galicia-Campus Terra (2022).OA e PRV foron financiadas por contratos predoutorais da Xunta de Galicia(Refs. ED481A-2020/119 e ED481A-2020/225, respectivamente). DT é financiada por un contrato predoutoral industrial da Xunta de Galicia (Ref. 06_IN606D_2022_2693134). SVR foi financiada por un contrato predoutoral FPU do Ministerio de Universidades (Ref. FPU18/00402).

Referencias

As referencias deste resumo pódense consultar empregando este código QR

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

R y lme4: Uniendo fuerzas para la Estimación en Áreas Pequeñas

Naomi Diz-Rosales[1], María José Lombardía[2] e Domingo Morales[3]

[1,2]Universidade da Coruña, CITIC, A Coruña, España.
[3]Universidad Miguel Hernández de Elche, IUI-CIO, Elche, España.

RESUMO

La Estimación en Áreas Pequeñas es una rama multidisciplinar de la estadística que asume el reto de obtener estimaciones precisas sobre variables de interés en dominios o áreas con tamaño muestral reducido o incluso nulo. Desde su origen, a finales de los 70, ha gozado de un desarrollo continuo, aplicándose a multitud de problemáticas. Ante este contexto, en los últimos años se han popularizado las aproximaciones metodológicas basadas en modelos mixtos. Éstos, permiten combinar diferentes fuentes de información, vinculando las observaciones de todas las áreas a través de los efectos fijos, al mismo tiempo que se incorpora la variabilidad existente entre los dominios definiendo efectos aleatorios que permiten que las áreas similares unan fuerzas incrementando el tamaño efectivo de la muestra. En este aumento de popularidad ha desempeñado y desempeña un papel clave lme4, el paquete de R referente en modelos mixtos. Esta poderosa herramienta para realizar análisis de modelos mixtos, lineales y generalizados, cuenta con numerosas ventajas, destacando su accesibilidad, constante actualización y amplia aplicabilidad. En esta ponencia, se presenta este paquete y su empleo en Estimación en Áreas Pequeñas ejemplificándolo con su aplicación a dos problemáticas de gran relevancia social como: la estimación de indicadores de pobreza por provincia y sexo en España y la estimación de indicadores de colapso asistencial en contextos pandémicos, como el vivenciado con la COVID-19. Si bien, por su gran versatilidad, el aprendizaje sobre el paquete es extensible a todos los ámbitos de la ciencia.

Palabras e frases chave: Estimación en Áreas Pequeñas, lme4, modelos mixtos, R.

1. INTRODUCIÓN

La Estimación en Áreas Pequeñas o *Small Area Estimation* (SAE) es una rama multidisciplinar de la estadística que objetiva la obtención de estimaciones precisas en áreas o dominios, geográficos o de otra índole, en las que el número de observaciones disponibles es muy reducido o incluso nulo. Desde su definición, con el estudio de Fay y Herriot [5], Battese, Harter y Fuller [2], Prasad y Rao [8] y Jiang y Lahiri [6], las investigaciones se han sucedido, aumentando significativamente en los últimos años impulsadas por los Objetivos de Desarrollo Sostenible (ODS), para garantizar el buen comportamiento de los indicadores globales desagregados por ingresos, sexo, edad, raza, ubicación geográfica, etc.

Como parte de este desarrollo destacan, especialmente, los modelos mixtos (MMs), que además de efectos fijos, incorporan efectos aleatorios. Para ello, los MMs tienen una estructura compleja multinivel o jerárquica a través de la cual se incorpora la posible dependencia existente entre variables que se observan en los mismo grupos. Esta propiedad es sumamente útil en el ámbito SAE, al premitir incrementar el número de observaciones efectivo modelando la variabilidad existente entre y dentro de las áreas o dominios. De este modo, si se asigna un mismo efecto aleatorio a áreas que compartan características similares, se está obteniendo un mayor número de observaciones

efectivo para realizar las estimaciones que si solamente se tomase en cuenta cada área de modo individual. Para una revisión detallada sobre esta metodología en particular, y sobre SAE en general, se pueden consultar los libros de Rao y Molina [9] y Morales, Esteban, Pérez y Hobza [7]. La comprensión del método es fácil si pensamos en el principio de "la unión hace la fuerza" y tenemos en mente la imagen de un pez pequeño indefenso ante un pez grande, pero valiente ante él cuando une fuerzas con sus iguales. La comunidad del software libre R no es ajena a este principio, y con un notable trabajo colaborativo, en la última década destaca la continua evolución, accesibilidad y optimización del paquete "lme4", referente en el estudio y aplicación de modelos mixtos.
El paquete lme4 en R es una herramienta poderosa para realizar análisis de modelos mixtos lineales y generalizados (LMM y GLMM, por sus siglas en inglés), pudiendo enfatizar, entre sus numerosas ventajas [1]:

- Flexibilidad: lme4 permite definir una amplia variedad de estructuras jerárquicas y anidadas en los datos, incorporando tanto efectos aleatorios como fijos, lo que lo hace adecuado para una amplia variedad de investigaciones. Además, cabe destacar que, no solamente se trabaja con LMMs, lme4 también admite GLMMs, lo que permite tratar, por ejemplo, variables de tipo conteo.
- Programación amigable: lme4 utiliza el mismo lenguaje de especificación de las fórmulas que otras funciones existentes en R para analizar modelos de regresión, como lm() y glm(), lo que facilita su empleo.
- Eficiencia computacional: lme4 utiliza algoritmos de optimización, adaptados al cómputo en paralelo, que pueden manejar conjuntos de datos grandes y modelos complejos.
- Diagnósticos y gráficos: lme4 proporciona herramientas de apoyo en la diagnosis del modelo, como gráficos de residuos y efectos aleatorios, lo que facilita la evaluación de la calidad del ajuste.
- Sentido de comunidad y desarrollo activo: lme4 cuenta con una sólida documentación y una comunidad activa y colaborativa, existiendo una amplia variedad de recursos y tutoriales, a la par que se actualiza y mejora continuamente.
- Compatibilidad con otros paquetes: las estructuras generadas con lme4 se reconocen por otros paquetes de R, esencial a efectos de generar mapas o gráficas informativas, obtener intervalos de confianza, realizar pruebas de hipótesis sobre los efectos aleatorios o, incluso, estimar el criterio de información de Akaike condicionado, métrica referente en SAE para la selección de modelos.

En conclusión, objetivamos dar a conocer este paquete y su empleo en SAE, ejemplificándolo con su aplicación a dos problemáticas de gran relevancia social, como la estimación de indicadores de pobreza por provincia y sexo en España [4] y la tasa de ocupación en unidades de cuidados intensivos a causa del COVID-19 [3]. El aprendizaje sobre el paquete será extensible a multitud de campos de la estadística, pues los modelos mixtos son ideales para modelar datos jerárquicos, como, por ejemplo, los obtenidos en estudios longitudinales con mediciones repetidas para los mismos sujetos o tratamientos, algo común en investigaciones en psicología, biología y otras disciplinas.

AGRADECEMENTOS

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación y la Agencia Estatal de Investigación del Gobierno de España a través del Fondo Europeo de Desarrollo Regional [PID2022-136878NB-I00, PID2020-113578RB-I00 y PRE2021-100857 a Naomi Diz-Rosales financiado por MCIN/AEI/10.13039/501100011033]; por la Conselleria d'Innovació, Universitats, Ciéncia I Societat Digital de la Generalitat Valenciana [Prometeo/2021/063]; por la Consellería de Cultura, Educación, Formación Profesional e Universidades de la Xunta de Galicia a través del Fondo Europeo de Desarrollo Regional [Grupos de Referencia Competitivos ED431C/2020/14, COV20/00604 and ED431G/2019/01]; y por el Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC) que está financiado por la Xunta de Galicia a través del convenio de colaboración entre la Consellería de Cultura, Educación, Formación Profesional e Universidades y las universidades gallegas para el refuerzo de los centros de investigación del Sistema Universitario de Galicia (CIGUS).

Referencias

[1] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P, Bauer, A., Krivitsky, P.N. (2023). Package lme4, version 1.1-34. Available at: `https://cran.r-project.org/web/packages/lme4/lme4.pdf`. (Accessed: September 2023).

[2] Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83, 28–36.

[3] Diz-Rosales, N., Lombardía, M.J., Morales, D. (2023). Modelling the COVID-19 ICU occupancy with area-level random regression coefficient Poisson models. In Pardo-Fernández, J.C, Rodríguez-Álvarez, M.X. (eds.): *Proceedings of the XIX Conferencia Española y VIII Encontro Iberoamericano de Biometría (CEB-EIB 2023)*, Vigo, España.

[4] Diz-Rosales, N., Lombardía, M.J., Morales, D. (2023, accepted for publication). Poverty mapping under area-level random regression coefficient Poisson models. *Journal of Survey Statistics and Methodology.* DOI: 10.1093/jssam/smad036.

[5] Fay, R. E., Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 269–277.

[6] Jiang, J., Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics* 53, 217–243.

[7] Morales, D., Esteban, M.D., Pérez, A., Hobza, T. (2021). *A course on small area estimation and mixed models. Methods, theory and applications in R.* Springer, Switzerland.

[8] Prasad, N.G.N., Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association* 85, 163–171.

[9] Rao, J.N.K., Molina, I.(2015). *Small area estimation. (2nd ed.).* Wiley, Hoboken.

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Series de Tempo con R

Manuel Febrero Bande[1]

[1]Facultade de Matemáticas, Universidade de Santiago de Compostela

RESUMO

O obxectivo desta palestra é recordar as ferramentas básicas para o tratamento de series de tempo dispoñibles no R. Estas ferramentas sufriron poucas modificacións no últimos anos aínda que si engadiron algunha novidade interesante. A palestra vaise centrar nas series de tempo regulares.

Palabras e frases chave: bases de datos, modelos ARIMA, GARCH

1. INTRODUCIÓN

Unha serie de tempo é unha secuencia ordeada de variables aleatorias $\{X_t\}_{t\in T}$ respecto a un índice t (que soe ser o tempo). Segundo a frecuencia de observación as series de tempo poden ser continuas ou discretas e, destas últimas, observadas a intervalos regulares ou irregulares. Con moita diferenza a series de tempo discretas observadas a intervalos regulares son as máis tratadas na literatura de series de tempo. Estas series coñécense como *series de tempo regulares*. O tratamento en `R` de este tipo de series involucra gardar, transformar e manipular tanto os valores da serie coma os seus índices temporais. Para tal fin, están creadas estruturas (obxectos) para os valores da series (`ts`, `timeSeries`) con diferentes propiedades e xeitos de gardar o índice temporal (ver paquetes `base`, `chron`, `lubridate`, `timeDate`). Na palestra farase un percorrido rápido polos distintos tipos de obxectos.

2. IMPORTACIÓN DE DATOS

Importar datos de series de tempo no `R` é usualmente un traballo recorrente na súa análise. Particularmente, as novidades neste punto corresponden a como importar directamente de follas web onde se crean táboas específicas ou mediante algún paquete que dispoña de funcións para a súa consulta e importación automática. Neste apartado destaca a libraría `quantmod` que permite consultar moitas bases de datos internacionais.

3. MODELOS ARIMA e GARCH

Os modelos máis populares para tratar series regulares con intervalo temporal maior que o diario son, sen dúbida, os modelos ARIMA (tamén coñecidos como modelos Box–Jenkins). Para estes modelos clásicos existen ferramentas no paquete `base` que funcionan bastante ben aínda que nos últimos anos colleron certo pulo algúns paquetes que ofrecen algunha funcionalidade máis ca básica (por exemplo o paquete `forecast`). As series con intervalo temporal menor ou igual que o diario soen estar vencelladas a sinais financeiras e adoitan a presentar o que se chama como *Volatilidade condicional*. Para este particular, deseñaronse os modelos GARCH e tamén no `R` temos librarías axeitadas como `rugarch` ou `fGarch`. Na palestra farase un repaso rápido que lle sirva a un usuario novel como punto de partida.

Referencias

[1] Chan, K.S. and Ripley, B. (2022) TSA: Time Series Analysis (1.3.1). http://cran.r-project.org/web/packages/TSA/.

[2] Galanos, A. (2023) rmgarch: Multivariate GARCH models (1.3-9). http://cran.r-project.org/web/packages/rmgarch/.

[3] Galanos, A. (2023) rugarch: Univariate GARCH models (1.5-1). http://cran.r-project.org/web/packages/rugarch/.

[4] Hyndman, R. (2023) forecast: Forecasting Functions for Time Series and Linear Models (8.21.1). http://cran.r-project.org/web/packages/forecast/.

[5] James, D. (2023) chron: Chronological Objects which can handle Dates and Times (2.3-61). http://cran.r-project.org/web/packages/chron/.

[6] Ryan, J.A. (2023) quantmod: Quantitative Financial Modelling Framework (0.4.25). http://cran.r-project.org/web/packages/quantmod/.

[7] Spinou, V. (2023) lubridate: Make Dealing with Dates a Little Easier (1.9.3). http://cran.r-project.org/web/packages/lubridate/.

[8] Wuertz, D. (2023) timeDate: Rmetrics - Chronological and Calendar Objects (4022.108). http://cran.r-project.org/web/packages/timeDate/.

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Uso da linguaxe R como "glue language" para procesamento de expedientes administrativos

Marcos Fernández Arias[1]

[1] Axencia de Modernización Tecnolóxica de Galicia, Xunta de Galicia

RESUMO

Breve exemplo práctico do uso da linguaxe R para automatización de procesos administrativos.

Palabras e frases chave: glue language, ETL,

1. INTRODUCIÓN

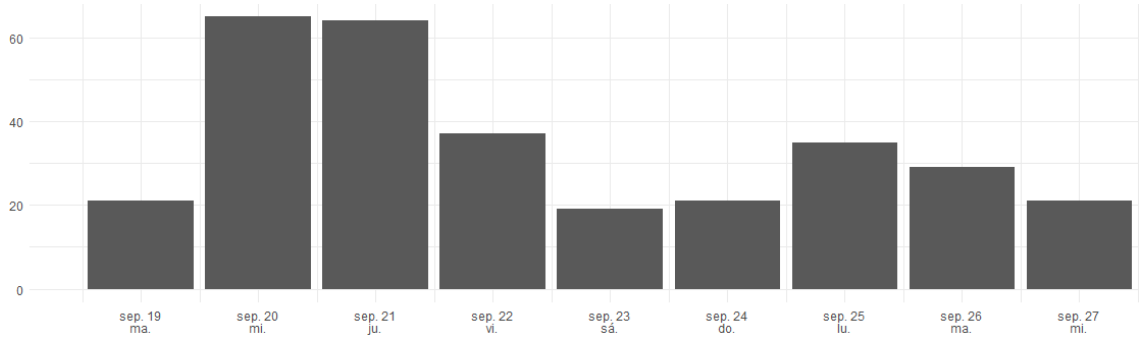
Unha linguaxe pegamento, do inglés "glue language", é en informática unha linguaxe de programación usada para unir compoñentes de software (intercambiar datos entre eles, chamar ás súas funcións, etc.) que en principio non foron deseñados expresamente para interactuar entre eles.

R pode utilizarse como solución para integrar sistemas de software dispares e descentralizados con recursos relacionados. Permite as operacións integradas de diferentes sistemas, independentemente da súa orixe e desenvolvedor.

R permite a colaboración de aplicacións locais e servizos web e automatizar os seus procesos.
Por exemplo, unha organización pode ter diferentes sistemas de software para a entrada de solicitudes, a tramitación de expedientes e o envío de notificacións.
Mediante un "glue language" é posible "pegar" todos eles para que poidan comunicarse e interactuar, funcionando como un sistema integrado.

2. BREVE EXEMPLO DO USO DA LINGUAXE R PARA O PROCESAMENTO DE EXPEDIENTES

Esplicarmos brevemente cómo utilizamos R na Xunta de Galicia para automatizar tarefas complementarias no procesamento de expedientes administrativos.
E, neste caso, para os expedientes de solicitudes de prazas en balnearios.
https://sede.xunta.gal/detalle-procedemento?codtram=BS607A




```
502   # gráfica de volumen de entrada de solicitudes en Sede
503   solic_0 %>%
504     filter(!str_detect(numero_rexistro, str_c("^", ano)))
505   solic_0 %>%
506     transmute(data_solicitude = as_date(data_solicitude)) %>%
507     count(data_solicitude) %>%
508     ggplot(aes(x = data_solicitude, y = n)) +
509     geom_col() +
510     scale_x_date(
511       breaks = "1 days",
512       date_labels = "%b %d\n%a"
513     ) +
514     labs(x = NULL, y = NULL) +
515     theme_minimal()
516 }
517
```

```r

# turnos ====
turnos_ofertados <- dbGetQuery(conn, glue(
  "SELECT turnos.id turno_id,
centro_id, fecha_inicio, fecha_fin,
centros.denominacion, centros.direccion
from turnos
inner join centros on turnos.centro_id = centros.id
where turnos.convocatoria_id = {CONVOCATORIA_ID};")) %>%
  as_tibble() %>%
  mutate(
    across(ends_with("_id"), as.integer)
  )
turnos_ofertados




# ********************************************************************************
# ********************************************************************************
fname <- "origen/BS607ASolPresent2023 (2023-09-28.ods"
stopifnot(fs::file_exists(fname))
expected_colnames <- c(
  "codigo", "numero_rexistro", "nome", "nif", "email_solicitante",
  "telefono_1_solicitante", "telefono_2_solicitante", "tipo_via",
  "nome_via", "numero", "bloque", "andar", "porta", "localidade",
  "codigo_postal", "provincia", "concello", "parroquia", "lugar",
  "data_solicitude", "nome_representante", "nif_representante",
  "email_notificacion", "telefono_notificacion", "estado", "tarxeta_sanitaria",
  "data_nacemento_solicitante", "contia_mensual_solicitante",
  # "pension_solicitante", "outros_solicitante",
  "indicar_solicitante", "nome_fillo", "dni_fillo",
  "data_nacemento_fillo", "tarxeta_sanitaria_fillo", "porcentaxe_discap",
  "nome_acompanante", "dni_acompanante", "data_nacemento_acompanante",
  "tarxeta_sanitaria_acompanante", "contia_mensual_acompanante",
  "pension_acompanante", "outros_acompanante", "indicar_acompanante",
  "destino_1", "fecha_1", "destino_2", "fecha_2", "destino_3", "fecha_3",
  "vacantes_en_otros_turnos"
)
# setdiff(expected_colnames, colnames(solic_0))
# setdiff(colnames(solic_0), expected_colnames)

solic_0 <- readODS::read_ods(fname, verbose = TRUE) %>%
  janitor::clean_names() %>%
  rename(
    fecha_1 = fecha_45, fecha_2 = fecha_47, fecha_3 = fecha_49,
  ) %>%
  select(-procedente) %>%
  verify(names(.) == !!expected_colnames) %>%
  rename(
    representante_nif = nif_representante,
    representante_nombre = nome_representante
  ) %>%
  janitor::remove_empty("cols") %>%
  as_tibble() %>%
  verify(estado == "SOLICITUDE PRESENTADA") %>%
  #select(-estado) %>%
  verify(!is.na(numero_rexistro) & !duplicated(numero_rexistro)) %>%
  verify(vacantes_en_otros_turnos %in% c("Si", "No")) %>%
  mutate(
    observaciones = "",
    across(starts_with("data_nacemento"), dmy),
    across(starts_with("nome"),
           (\(x) sub("\\b(D[eo]l?)\\b", "\\L\\1", str_to_title(x), perl = TRUE))
    ),
    axe_discap = na_if(porcentaxe_disca
    r replace(nome, "^(.*)
```

X Xornada de Usuarios de R en Galicia

Santiago de Compostela, 18 de outubro do 2023

MESA REDONDA «DEZ ANOS CELEBRANDO A XORNADA DE USUARIOS DE R EN GALICIA»

Mª José Ginzo Villamayor[1], Miguel Rodríguez Muíños[2] e Rafael Rodríguez Gayoso[3]

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

[2] Dirección Xeral de Saúde Pública da Consellería de Sanidade (Xunta de Galicia)

[3] Asociación de usuarios de software libre da Terra de Melide

PREGUNTAS

1. Historia e Evolución:
   - Como evolucionou o software R dende os seus inicios ata a actualidade?
   - Que vantaxes ofrecía R nos seus comezos fronte a outras solucións estatísticas?

2. Aplicacións e usos:
   - En que campos ou disciplinas atopou R a súa maior aplicación e por que?
   - Como impactou R na investigación científica e na toma de decisións empresariais?

3. Funcionalidades e Ferramentas:
   - Cales son as principais bibliotecas ou paquetes de R que consideran indispensables e por que?
   - Como se compara R con outras aplicacións estatísticas en termos de funcionalidade e flexibilidade?

4. Educación e Aprendizaxe:
   - Que recursos recomendarían para aqueles que desexan aprender R dende cero?
   - Cales son os principais desafíos que enfrontan os principiantes ao aprender R e como poden superalos?

5. Comunidade e Contribucións:
   - Como contribuíu a comunidade de R ao desenvolvemento e mellora do software?
   - Que papel xogan as conferencias e encontros de usuarios de R no fortalecemento da comunidade?

6. Integración e Compatibilidade:
   - Como se integra R con outras ferramentas e plataformas, como Python, SQL ou ferramentas de big data?
   - Que solucións existen para mellorar a eficiencia e velocidade de R ao manexar grandes conxuntos de datos?

7. Futuro e Desenvolvemento:
   - Que innovacións ou melloras esperan ver en R nos próximos anos?
   - Como ven o futuro de R fronte á crecente popularidade doutras linguaxes de programación en análise de datos?

8. Casos de Éxito:
- Poderían compartir exemplos de proxectos ou investigacións onde R fose fundamental para obter resultados significativos?
- Como axudou R a transformar a forma en que se abordan problemas complexos nos seus respectivos campos?

9. Aspectos Técnicos:
- Que consellos darían para optimizar o rendemento de R en proxectos de gran envergadura?
- Como manexan as limitacións de memoria e procesamento ao traballar con R?

10. Aspectos Éticos e Sociais:
- Dado que R é software libre, como impacta isto na democratización da análise de datos e a ciencia?
- Que responsabilidades teñen os usuarios e desenvolvedores de R en canto á ética na análise de datos?

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Design e análise de experimentos com R

Ariel Levy[1], Eduardo Camilo da Silva[1], Marcus Antonio Cardoso Ramalho[1] , Mariana Marinho da Costa Lima Peixoto[1]

[1] PPGAd - Universidade Federal Fluminense

RESUMO

Este trabalho explora o uso do R como uma ferramenta eficaz no desenvolvimento, design e análise de experimentos científicos. O R simplifica a documentação e o compartilhamento de experimentos, possibilitando a reprodução por outros pesquisadores. A integração com ferramentas de documentação e editoração científica é destacada, bem como a disponibilidade de literatura e recursos no CRAN. Conclui-se que a seleção cuidadosa de ferramentas, considerando suas limitações, é essencial, e incentiva-se os pesquisadores a explorar o potencial do R no processo de design e análise de experimentos científicos.

Palavras chave: Experimentos, R, Design e análise de experimentos

1. INTRODUÇÃO

Um experimento é um tipo de pesquisa científica que consiste na investigação de um fenômeno pressuposto a partir da relação de duas ou mais variáveis que deve ser conduzido de forma sistemática testando as hipóteses propostas.

Uma característica dos experimentos é a necessidade de documentação de todas as etapas, do design aos resultados. A documentação normalmente requer o uso de ferramentas específicas para cada etapa, o que pode gerar problemas de compatibilidade entre as ferramentas e dificultar a reprodução dos experimentos.

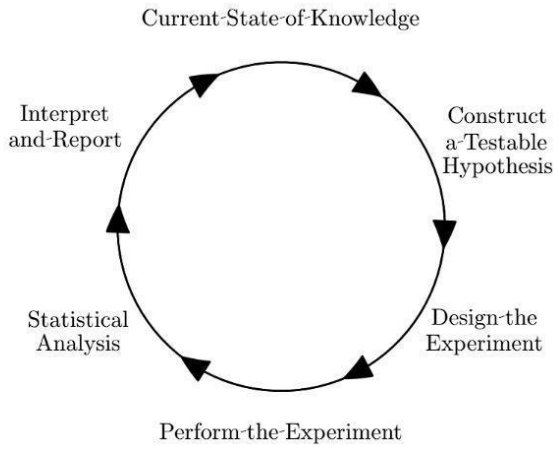


Figura 1 Adaptado de (Seltman, 2018)

Assim, todo experimento visa resolver uma questão de pesquisa que explica determinada relação e para tal precisa seguir um fluxo similar ao da figura 1.

As fases ali determinadas são facilmente compreendidas mas nem sempre de igualmente simples definição e execução. Nesta palestra pretendemos mostrar a utilização do R como veículo facilitador destas etapas.

2. Experimentos com R

Um possível fluxo de trabalho para elaboração de um experimento em R e sua documentação, em um arquivo QUARTO, irá requerer do pesquisador apenas o conhecimento básico de markdown e R. Interfaces como o VSCODE ou Rstudio permitem automatizar o processo de versionamento com o Github*. Assim, os arquivos e dados compartilhados através da ferramentas de versionamento podem ser reproduzidos por outros pesquisadores, gerando novos resultados ou versões, que serão incorporados ao documento original, ou não, promovendo um ciclo de pesquisa, aprendizado, inovação e desenvolvimento colaborativo

A utilização de ferramentas baseadas em markdown e Latex, voltadas para documentação e editoração científica como é o caso do QUARTO (Allaire et al., 2022) permite a produção de apresentações, artigos, sites e livros de forma padronizada e rápida. Uma vantagem em relação a outras ferramentas de documentação é a sua integração com a API do Zotero, o que permite o compartilhamento de referências bibliográficas entre os membros de um grupo de pesquisa e integração dessas referências no arquivo de bibliografia gerado pelo QUARTO. Através destes documentos ainda será possível integrar python, R e outras linguagens de programação.

A utilidade do R para design e análise de experimentos fica evidente ao verificarmos que é prolífica a literatura no tema, com obras inteiras dedicadas ao uso do R em experimentos científicos (Lawson, 2015). Em seu livro intitulado Design and Analysis of Experiments with R, Lawson (2015) apresenta uma série de exemplos e exercícios de experimentos científicos e suas respectivas análises usando o R.

A motivação para iniciar um design experimental pode acontecer de diversas formas, uma delas é através de uma observação, EDA (exploratory data analysis) em dados brutos provenientes de indicadores, ou por inquietações e questionamentos. Ainda na parte de exploração de dados e análises no R, existe a possibilidade de usar bibliotecas como o sos (Graves et al., 2023) e ctv (Zeileis & Hornik, 2023) para buscar no CRAN (Repositório de pacotes do R), funções e pacotes que atendam a necessidades específicas, diminuindo a necessidade de implementação novas funções.

O pacote sos recebe uma query como entrada e retorna uma lista de pacotes que atendam a demanda especificada, por exemplo, para buscar pacotes que tenham implementado a função rhoDCCA , podemos usar a seguinte query:

```
pkg <- sos::findFn("rhodcca")
```

```
found 9 matches
Downloaded 9 links in 4 packages.
```

```
pkg <- dplyr::glimpse(pkg)
```

```
Rows: 9
Columns: 10
$ Count       <dbl> 3, 3, 3, 3, 3, 3, 2, 2, 1
$ MaxScore    <dbl> 75, 75, 75, 45, 45, 45, 44, 44, 22
$ TotalScore  <dbl> 110, 110, 110, 86, 86, 86, 62, 62, 22
$ pkgLink     <chr>
"https://search.r-project.org/CRAN/refmans/DFA/html/00Inde…
$ Package     <chr> "DFA", "DFA", "DFA", "SlidingWindows",
"SlidingWindows", "…
$ Function    <chr> "rhoDCCA", "00Index", "Deltarho",
"rhodcca.SlidingWindows"…
$ Date        <dttm> 2023-07-11 10:10:07, 2023-07-11 10:10:07,
2023-07-11 10:10…
$ Score       <dbl> 75, 22, 13, 45, 23, 18, 44, 18, 22
$ Description <chr> "R: Detrended Cross-Correlation Coefficient
(rhoDCCA)", "R…
$ Link        <chr>
"https://search.r-project.org/CRAN/refmans/DFA/html/rhoDCC…
```

Já a função ctv retorna pacotes avaliados como relevantes pela comunidade para tarefas específicas de alguns campos de pesquisa. Para verificar pacotes relevantes para o campo de design de experimentos, podemos usar a seguinte query:

```
a <- ctv::ctv("ExperimentalDesign")
```

```
name: "ExperimentalDesign",
topic: "Design of Experiments (DoE) & Analysis of Experimental Data",
maintainer: "Ulrike Groemping, Tyler Morgan-Wall",
email: "ulrike.groemping@bht-berlin.de",
version: "2023-04-05",
source: https://github.com/cran-task-views/ExperimentalDesign/,
packagelist: [  105 items  ],
archived: "dfpk",
citation: [  1 item  ],
repository: https://cloud.r-project.org
```

Figura 2: Resultado da query ctv("ExperimentalDesign")

Os resultados da Figura 2 retornam que existem 105 pacotes avaliados e que o repositório dedicado a catalogar estes pacotes é mantido atualizado, porém, a maioria das bibliotecas é direcionada para experimentos em campos específicos da ciência, o que não invalida a possibilidade de aproveitar funcionalidades comuns.

O Taskview (Zeileis & Hornik, 2023) aponta que existem alguns pacotes generalistas para análise de experimentos, e que os restantes estão associados a áreas de ensaios clínicos, agricultura, indústria e outras ainda mais específicas.

Durante a Xornada teremos oportunidade de apresentar as funcionalidades de alguns destes pacotes generalistas utilizando o pacote ctv.

3. CONCLUSÃO

Embora os pacotes a serem apresentados se refere sejam generalistas, não será possível esgotar o tema durante o evento. Sendo possível que se encontrem soluções em outros pacotes. Uma seleção mais ampla pode ser encontrada Zeileis & Hornik (2023). Uma consideração importante é que o uso de pacotes, apesar de facilitar o processo de desenvolvimento da pesquisa deve vir acompanhado de rigor, pois, apesar das soluções estarem disponíveis, nem sempre elas são adequadas para o problema em questão, por isso é importante conhecer as ferramentas e suas limitações.

Referências

[1] Allaire, J. J., Teague, C., Xie, Y., & Dervieux, C. (2022). Quarto. Zenodo. https://zenodo.org/record/5960048

[2] Graves, S., Dorai-Raj, S., Francois, & Romain. (2023). sos: Search Contributed R Packages, Sort by Package. https://cran.r-project.org/web/packages/sos/index.html

[3] Lawson, J. (2015). Design and analysis of experiments with R. CRC Press.

[4] Seltman, H. J. (2018). Experimental Design and Analysis.

[5] Zeileis, A., & Hornik, K. (2023). ctv: CRAN Task Views. https://cran.r-project.org/web/packages/ctv/index.html

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

SISTEMAS DE RECOMENDACIÓN A PARTIR DE TÉCNICAS DE CLÚSTERING BASADAS EN LA ESTIMACIÓN TIPO NÚCLEO DE LA DENSIDAD

Lucía López López[1] y Paula Saavedra-Nieves[2]

[1] Universidad de Santiago de Compostela
[2] CITMAga, Universidad de Santiago de Compostela

RESUMEN

El análisis clúster se refiere a una extensa variedad de métodos utilizados para explorar conjuntos de datos con el fin de encontrar grupos de observaciones similares. Tradicionalmente, este objetivo se ha logrado mediante la evaluación de alguna medida de distancia o disimilitud entre los elementos. La estimación de regiones de elevada densidad de la distribución de probabilidad subyacente proporciona un enfoque alternativo introducido, por primera vez, en Hartigan (1975). A diferencia de otros algoritmos de clasificación, esta metodología no requiere establecer el número de grupos de antemano, y la flexibilidad de los estimadores no paramétricos de la densidad permite detectar clústers de forma arbitraria (ver Menardi y Azzalini, 2014). La implementación de estos algoritmos en el paquete pdfCluster de R (ver Azzalini y Menardi, 2014), nos ha permitido proponer un sistema de recomendación musical a partir de datos reales de la plataforma Spotify.

Palabras e frases chave: Estimación tipo núcleo. Regiones de elevada densidad. Sistemas de recomendación. Clústering.

Referencias

[1] Azzalini, A., Menardi, G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *Journal of Statistical Software*, 57(11), 1–26.

[2] Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

[3] Menardi, G., Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, 24(5), 753-767.

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

COLORINDO MANDALAS COM R: EXPLORANDO CORES E GRADIENTES EM CURVAS PLANAS

João Paulo M. Santos[1], Luciane F. Alcoforado [2]

[1]Academia da Força Aérea
[2]Academia da Força Aérea

RESUMO

Alguns métodos de coloração de figuras construídas com base em curvas paramétricas combinados com movimentos rígidos de rotação, translação e homotetias sucessivas são apresentados. A convergência de conceitos envolvendo matemática e programação por meio do software R proporcionou o progresso das investigações cuja metodologia básica é o processo de atribuição de cores em formas denominadas sequencial e aleatória, respectivamente. A discussão deste resumo evolui a partir da descrição da capacidade de coloração utilizando apenas o R básico e foca nas cores do modelo RGB e nos mapas gradientes. Em seguida, o método empregado na referência básica é modificado para proporcionar a expansão da capacidade de construções. Os resultados apontam para um novo conjunto de possibilidades de construções. Também indicam potenciais usos da linguagem R envolvendo o desenvolvimento de figuras complexas com características artísticas ou aplicações no ensino.

Palabras e frases chave: Mandalas. Coloração. Linguagem R. Curvas Planas. Geometria. Visualização. Matemática.

1. INTRODUCÃO

A convergência de conceitos que envolvem mandalas, matemática e programação é um processo de investigação que iniciou a partir da concepção inicial de gerar mandalas por meio da linguagem de programação R, possibilitando um campo de pesquisa multifacetado. A partir da formação de um grupo de estudos composto por docentes especializados na linguagem R e em geometria com curvas planas, atingimos o propósito coletivo de conceber funções geradoras de mandalas utilizando equações paramétricas de curvas clássicas.

Durante esse percurso, houve amplas discussões em torno das vantagens do R em relação ao Geogebra (https://www.geogebra.org/), especificamente no contexto dessas construções. Realizamos testes iniciais com ambos os programas, e a robustez do R o tornou a escolha preferencial para a execução desta pesquisa. O resultado inicial desse esforço conjunto foi publicado no livro Mandalas, Curvas Planas e Visualização com R (Alcoforado *et al.*, 2023). Além disso, utilizar o R no presente contexto adiciona um potencial de exploração tanto do ponto de vista computacional quanto da integração com outras áreas como a estatística.

O objetivo deste artigo é mostrar algumas possibilidades para os métodos de coloração de mandalas construídas na linguagem R, utilizando curvas planas e movimentos rígidos no plano, de rotação, translação ou homotetias. Os métodos de coloração aqui propostos utilizam as cores disponíveis no pacote básico do R, em específico aquelas disponíveis em *colors*() ou nos mapas gradientes do tipo *heat.colors*() ou *rainbow*(), por exemplo. Isso elimina a dependência de outros pacotes e proporciona uma abordagem mais simplificada e acessível. Os resultados, por sua vez, são promissores e ressaltam o potencial a ser explorado.

2. APLICAÇÃO DO MODELO RGB EM MANDALAS

Para a visualização da figura construída por meio de procedimento de rotação, translação ou homotetias é interessante aplicação de um processo de coloração. Neste sentido, é necessário decidir a coloração de fundo e a coloração de pontos específicos da construção. A composição resultante é dependente, entre outros fatores, da forma de coloração e das cores que estão sendo utilizadas no processo.
Aqui o modelo modelo RGB, o qual utiliza escalas de vermelho (*Red*), verde (*Green*) e azul (*Blue*) foi utilizado. A utilização é simplificada pela disponibilidade de um conjunto de 657 cores por meio da função $colors()$. Estas podem ser acessadas diretamente e sem necessidade de instalação de pacotes. Também estão disponíveis algumas paletas do tipo gradientes no R básico. Uma discussão geral sobre cores no R pode ser encontrado, por exemplo, em (Frazier, 2020). Cores ou modelos adicionais podem ser viabilizados por meio de pacotes, mas a utilização de cores do R básico mostra que há um potencial para ser explorado em termos do vasto número de possibilidades. A complexidade dos elementos obtidos pode ser visualizada na Figura 1.

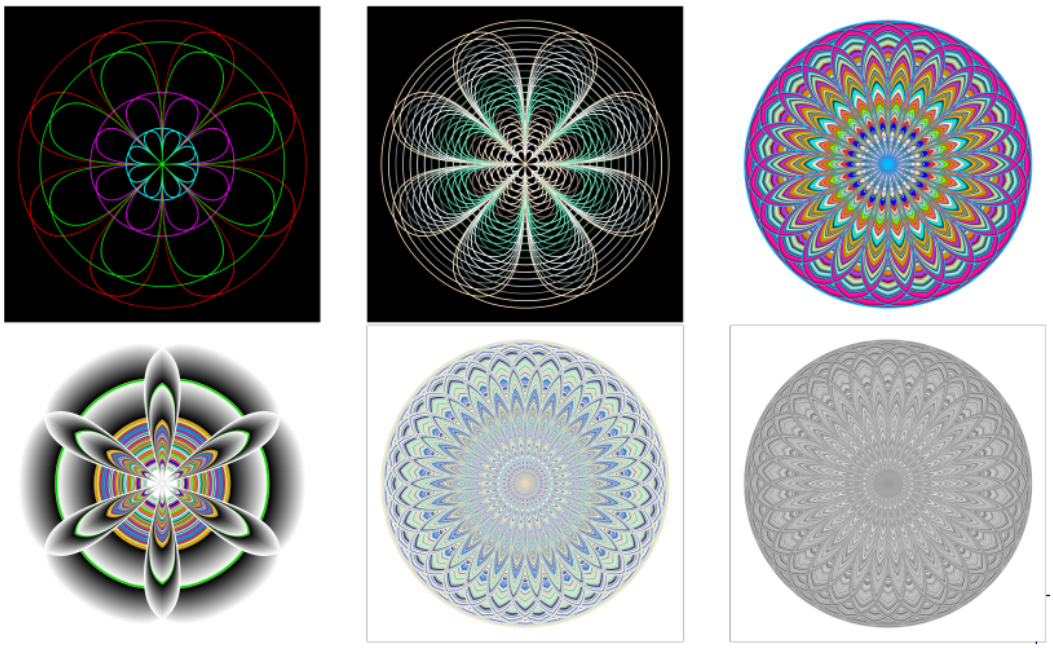


Figura 1: Figuras geradas com método de coloração sequencial e aleatório a partir de uma paleta p de cores disponíveis no R. Cores idênticas são atribuídas aos pontos de uma homotetia.

Todas as figuras em Figura 1 são resultantes do mesmo processo de composição de rotações da lemniscata de Gerono seguidas de homotetias. Lendo-se da esquerda para a direita, na primeira figura, o número de rotações, o número de homotetias e as cores utilizadas podem ser determinadas por inspeção. Na segunda, o número de rotações e homotetias ainda podem ser obtidos por inspeção, mas a paleta de cores não é mais evidente devido ao processo de sobreposição e interação com o plano de fundo. As figuras da terceira e quarta posições da sequência utilizam uma paleta específica de cores no intervalo de cores disponíveis com para um certo número de homotetias sendo o vetor de cores de mesmo comprimento do vetor de homotetias. As figuras finais da quinta e sexta posição ilustram o resultado de uma escolha inicial de cores cuja coloração resulta de um processo de seleção aleatória e com reposição dentre as cores escolhidas.
O processo de construção que fornece os resultados na Figura 1 segue os delineamentos utilizados e detalhados em (Alcoforado *et al.*, 2023) e podem ser resumidos em:

i.) Escolher as curvas ou figuras geométricas;

ii.) Aplicação de transformações geométricas;

iii.) Realizar a escolha do modelo de cores, escolha da paleta de cores e especificação do padrão a ser utilizado;

iv.) Composição de uma ou mais aplicações das transformações geométricas para obter um elemento único.

v.) Finalmente, a figura gerada é visualizada por meio do **ggplot2** ().

Os métodos de coloração apresentados em (Alcoforado *et al.*, 2023) consistem na exploração de aplicações sequenciais ou aleatórias da paleta de cores. Dada uma paleta $p = [p_1, \cdots, p_n]$ de cores, o método sequencial consiste na atribuição da cor p_k para a composição k; o método aleatório consiste em uma amostra aleatória, com reposição, p_a das cores em p. As Figuras 2 e 3 mostram os resultados das modificações no método de coloração utilizados em Figura 1.
A modificação no método sequencial e aleatório é a atribuição das cores ao vetor de pontos da composição. Nesse contexto, o método sequencial e aleatório ainda podem ser aplicados, desde que as alterações necessárias para contabilizar a quantidade de pontos na composição seja feita. Em termos mais detalhados, dado uma composição c com N pontos e uma paleta com $p = [p_1, \cdots, p_n]$, o vetor de cores é composto da escolha sequencial de p, na ordem estrita de ocorrência, ou da escolha aleatória e com reposição de cores em p de tal forma que o vetor de cores seja exatamente do comprimento do vetor de pontos da composição.

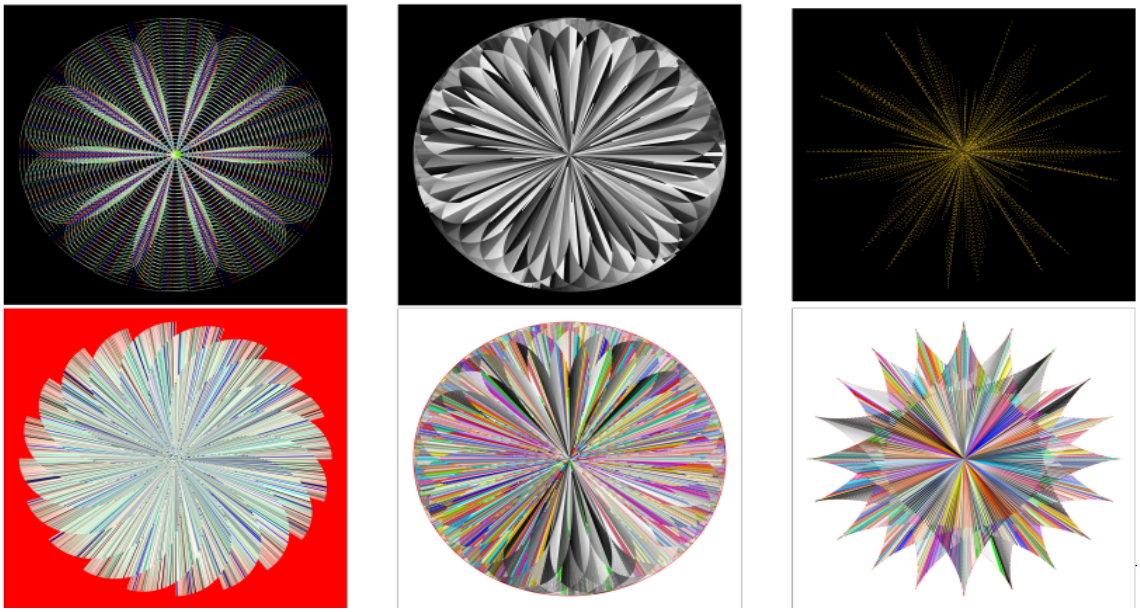


Figura 2: Figuras geradas com método de coloração sequencial e aleatório aos pontos de uma mesma homotetia a partir de uma paleta p de cores disponíveis no R.

A Figura 2 mostra alguns resultados: i.) as duas primeiras figuras resultam do processo sequencial com $p = colors()[1:50]$ e $p = colors()[154:250]$; ii.) a terceira é uma seleção aleatória e com reposição de $p = colors()[142:162]$ seguida de processo sequencial; iii.) a quarta é uma seleção aleatória e com reposição de $p = colors()[1:26]$ iv.) as duas últimas resultam do processo sequencial para $p = colors()$ com diferenciação entre as curvas planas utilizadas.
A aplicação da modificação anterior no contexto dos mapas gradientes fornecem os resultados mostrados na Figura 3. Neste caso, apenas os mapas $heat.colors()$ e $rainbow()$ foram explorados. A primeira figura utiliza todas as cores do mapa de cores $heat.colors()$; a segunda e terceira figuras utiliza $heat.colors(k)$, para $k = 100, 1000$ seguidas de uma escolha sequencial; a quarta figura utiliza todas as cores do mapa de cores $rainbow$; a quinta e sexta figuras utilizam métodos sequencial e aleatório, respectivamente.

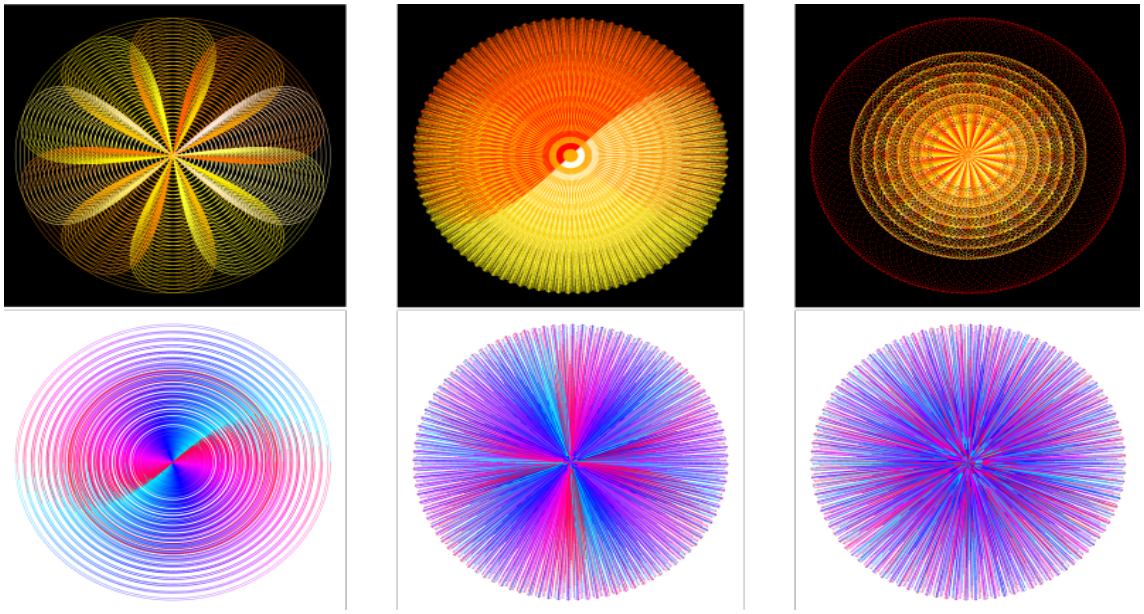


Figura 3: Figuras geradas com método de coloração sequencial e aleatório aos pontos de uma mesma homotetia a partir de uma paleta p de cores gradientes disponíveis no R.

3. CONCLUSÕES

O artigo explora a capacidade do R para produção de figuras com simetria radial. Os resultados seguem o processo delineado previamente e insere uma nova perspectiva com vistas a expandir o potencial de possibilidades. Os métodos, denominados sequencial e aleatório, no qual as cores de uma paleta são empregadas em todos os pontos de uma dada homotetia foram modificados para acomodar a coloração dos pontos na homotetia. Neste caso, as modificações inseridas resultam em outros distintos padrões de cores devido às diferentes possibilidades. Apesar disso, a aplicação está restrita às cores disponíveis na função *colors*() e alguns mapas gradientes. Tal restrição visa mostrar que há potencial a ser explorado apenas com o R básico aliado ao **ggplot2** (Whickhan, 2016). Esses resultados exploratórios mostram um potencial adicional à ser explorado; também é necessário observar que curvas planas adotadas, em geral, compõem uma pequena parcela das possibilidades anotadas na referência básica.

Referencias

[1] Alcoforado, L.F., Santos, J.P.M., Lima, M.V.A., Jesus, A.F., Linares, J. L. (2023). **Mandalas, curvas clássicas e visualização com R**. Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos. DOI: https://doi.org/10.11606/9786587023335 Disponível em aqui! . Acesso em 21 Set. 2023.

[2] Frazier, M. R Color CheatSheets (2020). **National Center for Ecological Analysis and Synthesis-NCEAS**. Disponível em: aqui!.

[3] R Core Team. (2021). R: A Language and Environment for Statistical Computing (Manual). **R Foundation for Statistical Computing**. https://www.R-project.org/.

[4] Wickham, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.

X Xornada de Usuarios de R en Galicia

Santiago de Compostela, 18 de outubro do 2023

A difusión de estatísticas públicas coas ferramentas que ofrece R

María Martín Vila[1], Antonio Albo Díaz[1] e Mª Esther López Vizcaíno[1]

[1] Instituto Galego de Estatística

RESUMO

A necesidade de difundir a información estatística de xeito sistemático e en formatos reproducibles pero á vez ofrecer produtos accesibles para toda a poboación lévanos a diversificar a difusión e mesturar táboas multidimensionais con produtos específicos máis elaborados. É neste punto onde R toma unha notable presenza posto que permite a través da xeración de informes con R- Markdown e coas aplicacións R-Shiny automatizar o traballo diario e presentalo en formatos axeitados.

Palabras e frases chave: R-Markdown, R-Shiny, estatística pública, difusión

1. INTRODUCIÓN

O Instituto Galego de Estatística (IGE) é un organismo autónomo da Xunta de Galicia creado no ano 1988 e que se rexe basicamente pola Lei 9/1988 de Estatística de Galicia. Na súa misión de promover o desenvolvemento do sistema estatístico da Comunidade Autónoma debe prestar servizos de recompilación e difusión da documentación estatística dispoñible, desenvolver bases de datos de interese público, analizar as necesidades e a evolución da demanda de estatísticas e asegurar a súa difusión.
Na consecución destes obxectivos seguimos a traballar para aumentar a interacción cos usuarios, de xeito que poidan conseguir os datos que precisen de forma sinxela, ofrecéndolles a oportunidade de consultar información a medida e de construír as súas propias táboas. Isto implica a xeración dinámica de páxinas e a disposición de bases e bancos de datos específicos de difusión que se poden interrogar en liña.
A nova web e os produtos específicos de difusión que incorpora tentan responder ás necesidades que teñen os cidadáns en termos numéricos no eido socioeconómico, contribuíndo así, aínda que dun xeito modesto, ao desenvolvemento da sociedade da información dende un punto de vista estatístico.

2. INFORMES ATOMATIZADOS E APLICACIÓNS DE DIFUSIÓN CON R-SHINY

No IGE emprégase R-Markdown para a xeración de informes automáticos tanto de corte conxuntural como estrutural. As estatísticas mensuais precisan dun método de actualización áxil de xeito que, feitas as actualizacións das táboas de datos correspondentes, o resumo se actualice executando un código escrito en R-Markdown que captura os datos necesarios da base de datos. Estes resumos de resultados son un compendio de táboas, gráficos e mapas xunto con texto explicativo da estatística en cuestión. Están programados con parámetros, que xunto cos demais elementos gráficos, se embeben no texto explicativo para unha maior automatización.

Para facilitar o traballo a todo o persoal do IGE elaborouse unha plantilla de R-Markdown e programáronse funcións que capturan non só datos, se non tamén metados da base de datos de MySql. Así, o proceso de lectura dos datos para a xeración de informes convértese nunha ou varias sentenzas de código.
Estes datos ofrécense ao cidadán en bruto xunto con explicacións, pero tamén como táboas xa definidas, gráficos ou mapas. Para isto contamos cunha serie de funcións: Make_Graphs_Bar, Make_Tables, Make_Choropleth,... que permiten transformar os data.frame ou data.table en táboas e gráficos embebidos no informe, cunha apariencia e formatos xa predefinidos e acorde coas directrices de difusión establecidas.

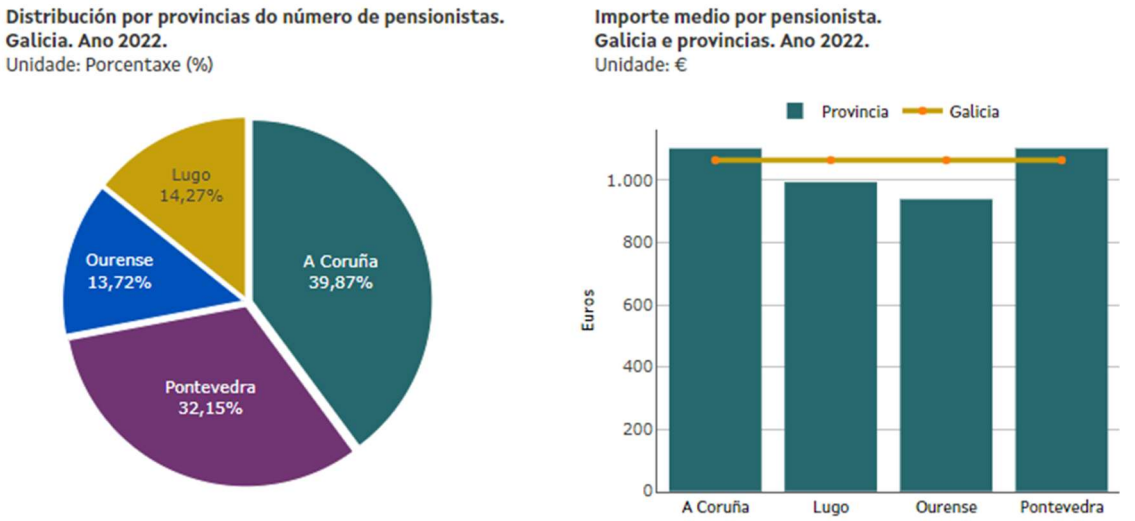


Figura 1: Gráfico de sectores e gráfico de barras da operación Pensións da Seguridade Social por concello de residencia da persoa pensionista

O emprego dunha plantilla e funcións auxiliares comúns axiliza de xeito importante o maquetado do informe tanto en estrutura, inclúese un apartado de índice e outro de definicións, descarga de datos e fontes de información, como de estilo, posto que tanto a letra como as distintas paletas de cor xa veñen predefinidas.
Ademais dos resumos de resultados, o IGE incorpora na súa páxina web aplicacións web dinámicas que permiten a interacción do usuario para consultar a información desexada. A filosofía deste tipo de aplicacións foi adaptándose á evolución da web e evolucionou desde aplicacións que aparecían como páxinas web "alleas" a, na última versión programada, aplicacións embebidas na web. Estas aplicacións permiten seleccionar por medio de despregables e/ou pestanas períodos ou variables. Son exemplos destas aplicacións para difusións específicas de información o Panorama dos sete grandes concellos, o Marco input-output ou os Indicadores de desenvolvemento sostible, entre outros.
A nova versión da web incorpora unhas aplicacións Shiny embebidas para cada unha das operacións do IGE. Estas aplicacións mostran, dun xeito moi visual e rápido, os conceptos e datos importantes da estatística que queremos consultar. As aplicacións Shiny de cada unha das operacións do IGE teñen todas a mesma estrutura e combinan catro recadros coa información máis relevante con dúas representacións gráficas, gráficos ou mapas, segundo resulte máis adecuado para a propia operación estatística.

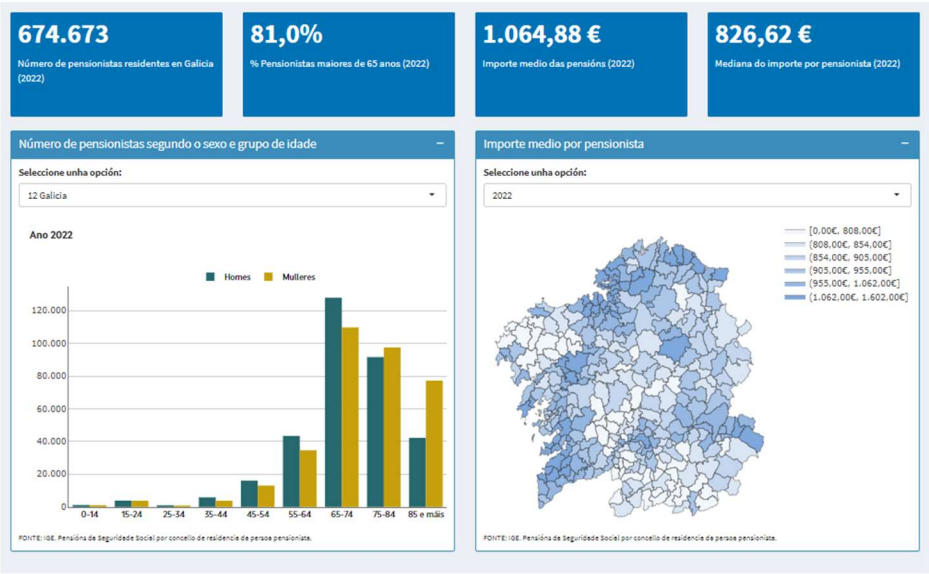


Figura 2: Shiny explicativo da operación Pensións da Seguridade Social por concello de residencia da persoa pensionista

Estas aplicacións actualízanse simultaneamente aos datos da actividade posto que len a información da mesma base de datos en MYSql, non obstante sérvense dunha base de datos auxiliar onde se almacena a configuración de cada unha das actividades, como poden ser os texto en distintos idiomas, a orixe de datos etc. Esta base de datos auxiliar contén tamén instrucións en R necesarias para o preprocesado dos datos que se implementarán nas aplicacións Shiny, principalmente nos gráficos.
A primeira vez que se consulta a aplicación Shiny para unha actividade concreta realízase a carga da configuración, que se garda en caché para servir ás seguintes peticións. Isto dota ás aplicacións de maior rapidez de carga, necesaria por seren consultadas de xeito recorrente e por múltiples usuarios.

Referencias

[1] 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/.

[2] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny.

[3] JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2021). rmarkdown: Dynamic Documents for R. R package version 2.7. URL https://rmarkdown.rstudio.com..

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Detección de anomalías usando el paquete `qcr`

Salvador Naya[1], Javier Tarrío-Saavedra[1], Miguel Flores[2] e Rubén Fernández–Casal[3]

[1]Grupo MODES. CITIC. Departamento de Matemáticas. Escola Poltécnica de Enxañaría de Ferrol, Ferrol, Universidade da Coruña.
[2]Departamento de Matemática. Grupo MODES. Facultad de Ciencias, Escuela Politécnica Nacional, Ladrón de Guevara E11–253, Quito, 17–01–2759, Pichincha, Ecuador.
[3]Grupo MODES. CITIC. Departamento de Matemáticas. Facultade de Informática, Campus de Elvi na, Universidade da Coruña.

RESUMO

A detección de anomalías é un proceso importante no control da calidade, xa que permite identificar patróns pouco habituais ou atípicos que poden indicar problemas ou eventos significativos. No contexto do software `R`, existen diversos paquetes que se poden utilizar para detectar anomalías nos conxuntos de datos, como pode ser o paquete `AnomalyDetection`. Neste traballo propoñemos o paquete qcr como alternativa para datos complexos expondo exemplos de aplicación en contextos de control de tráfico do Canal de Panamá ou en Eficiencia Enerxética.

Palabras e frases chave: Anomalías, Control da Calidade, `qcr`.

1. INTRODUCIÓN

No contexto do software `R` libre, existen varios paquetes e técnicas que se poden utilizar para detectar anomalías. Algúns dos máis usados son os seguintes:
1.- `AnomalyDetection`: este paquete ofrece funcións para detectar anomalías en series temporais univariantes e multivariantes. Utiliza un enfoque baseado na descomposición STL (descomposición estacional e tendencias mediante Loess).
2.- `autoencoder`: este paquete implementa redes neuronais de autoencoder para a detecciíon de anomalías en datos sen etiquetar. Os autocodificadores son modelos de aprendizaxe profunda que poden aprender representacións eficientes de datos e detectar anomalías en función da capacidade de reconstrución do modelo.
3.- `IsolationForest`: este paquete implementa o algoritmo Isolation Forest, que é un método non supervisado para a detección de anomalías. O algoritmo constrúe árbores de decisión aleatoria e utiliza a profundidade das árbores para medir a rareza das instancias.
4.- `stats`: implementa algoritmos de clasificación non supervisada como o método k-means que, aínda que se usa principalmente para identificar agrupacións, pode tamen ser útil con contexto de detectar anomalías.
Como alternativa propoñemos usar o paquete `qcr` [1], Quality Control Review que, aparte de implementar as principais ferramentas univariantes e multivariantes para o control estatístico de procesos e a análise da sua capacidade, tamén da a posibilidade de aplicar gráficos de control de tipo non paramétrico baseados no concepto de profundidade de datos, ademáis de proporcionar alternativas de gráficos de control para datos funcionais [2]. A continuación móstrase unha moi breve descrición das principais alternativas non paramétricas implementadas no paquete `qcr`, construidas a partires do cálculo da profuncidade de datos simplicial, da profundidade de Mahalanobis, da correspondente á máxima verosimilitude, a proposta por Tukey e, tamén, a de proxeccións aleatorias:

- Estatístico de rangos, r, alternativa aos gráficos de medidas individuais: $r_{G_m}(y) = \frac{\#\{D_{G_m}(Y_j) \leq D_{G_m}(y), j=1,\ldots,m\}}{m}$.
- Estatístico Q, alternativa ao gráfico de control Shewhart, $\bar{x}$: $Q(G_m, F_n) = \frac{1}{n}\sum_{i=1}^{n} r_{G_m}(X_i)$.
- Estatístico S, alternativa ao gráfico de control con memoria CUSUM: $S_n(G_m) = \sum_{i=1}^{n}\left(r_{G_m}(X_i) - \frac{1}{2}\right)$. Sendo $CL = 0$ e $LCL = -Z_\alpha\sqrt{n^2\frac{\left(\frac{1}{m}+\frac{1}{n}\right)}{12}}$.

Nas expresións anteriores, y é unha nova observación multivariante; Y_j, con $j = 1, \ldots, m$, representa a mostra de calibrado ou retrospectiva; $D_{G_m}(y)$ é a profundidade de y con respecto á distribución de calibrado, G_m; mentres que F_m é a distribución dunha mostra a monitorizar.
Actualmente estanse a incorporar novas funcións que permitan a detección de anomalías desde unha perspectiva da Aprendizaxe Máquina e a Minaría de Datos, en concreto para a aplicación do método Local Correlation Integral (LOCI) con datos multivariantes ou funcionais [3, 4]. Na seguintes sección se describe como usar este paquete e finalmente se comenta a aplicación a datos reais.

2. PAQUETE QCR

O paquete qcr en R ofrece ferramentas para o control de calidade estatístico, incluíndo gráficos de control univariantes e multivariantes, análise de capacidade de procesos e de estudos interlaboratorios.

3. APLICACIÓNS O CONTROL NO CANAL DE PANAMÁ

O transporte que se fai de buques na Canle de Panamá é excepcional e require de medidas de control onde os gráficos tipo Shewhart son unha interesante alternativa. Neste caso presentamos o control feito da variable tempo en tránsito a través de cada unha das esclusas do Canal de Panamá Expandido, tomado nos primeiros 42 meses de funcionamento das novas esclusas [5].

4. ANOMALÍAS EN EFICIENCIA ENERXÉTICA

A aplicación en eficiencia enerxética está ligada o desenvolvemento de tecnoloxías de IoT para o control e seguimento continuo das instalacións enerxéticas, incluídos os sistemas de climatización, co obxectivo de optimizar os recursos, aumentar a eficiencia e manter o confort térmico. O emprego de ferramentas de control de calidade que permitan a detección automática de anomalías resulta de gran interese. Esta tarefa proponse realizala mediante o control do proceso co paquete `qcr` [1]. Concretamente a idea parte das variables CTQ son monitorizadas continuamente con respecto ao tempo, e está baseada no traballo publicado de Barbeito et al. [6].
Igual de importante como o control e deteccóin de anomalías nun sistema HVAC, é frecuentemente o estudo da capacidade do sistema para cumprir as especificacións, para este fin o emprego do paquete `qcr` da a posibilidade do uso de indicadores non paramétricos que permiten verificar as especificacións de variables de interese como a temperatura ou o consumo.
A continuación se amosa un exemplo de aplicación da librería qcr, en concreto dos gráficos non paramétricos de rangos, para o control da eficiencia enerxética dunha tenda de roupa nun centro comercial de Panamá. Temos dúas variables críticas para a calidade do sistema, o consumo diario en climatización e o consumo en iluminación. As variables non son normais nin autocorreladas, polo que se poden aplicar gráficos de control non paramétricos. Pártese dunha mostra de calibrado con datos correspondentes aos luns, martes, mércores, xoves e venres. Os sábados ábrese unha hora máis que os luns-venres e dúas horas máis que os domingos. Polo tanto, detéctanse como alarmas os domingos (baixo consumo), as verdadeiras averías e paradas, ademáis dos sábados (alto consumo).

```
R>x <- as.matrix(Shop[c(44:dim(Shop)[1]),c(3,8)])
R>G <- as.matrix(Shop.week[c(1:30),c(3,8)])
R>data.npqcd <- npqcd(x, G)
R>res.npqcs <- npqcs.r(data.npqcd, method = 'Tukey', alpha = 0.0028)
R>plot(res.npqcs, title = 'r Control Chart')
```

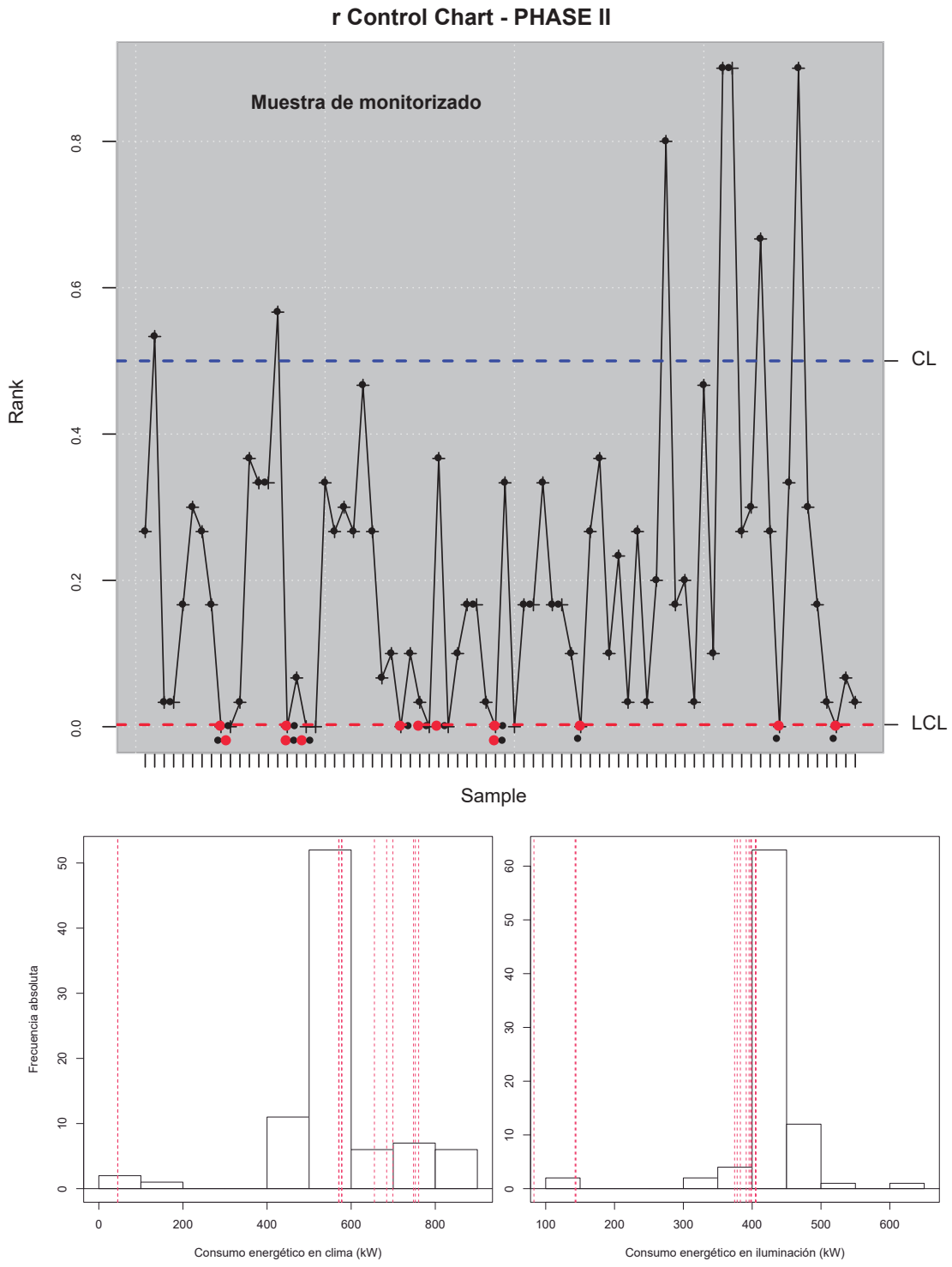


Figura 1: Gráfico de control r e histogramas do consumo en climatización e iluminación, cos días anómalos marcados.

5. CONCLUSIÓNS

O empleo do paquete qcr resulta unha alternativa para a detección de anomalías no contexto de datos complexos e usando conceptos propios do control estatístico da calidade.

AGRADECEMENTOS

Este estudo contou co apoio do Ministerio de Ciencia e Innovación coa subvención PID 2020-113578RB-100, ademais da Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 e Centro de Investigación do Sistema Universitario de Galicia ED431G2019/01).

Referencias

[1] Flores, M., R. Fernández-Casal, S. Naya, and J. Tarrío-Saavedra. (2022). qcr: Quality Control Review. `https://CRAN.R-project.org/package=qcr`.

[2] Flores, M., S. Naya, R. Fernández-Casal, S. Zaragoza, P. Raña, and J. Tarrío-Saavedra. (2020). Constructing a Control Chart Using Functional Data. Mathematics 8 (1): 58.

[3] Tobar, A., S. Flores M.and Castillo-Páez, S. Naya, S. Zaragoza, and J. Tarrío-Saavedra. (2023). Bootstrap-LOCI Data Mining Methodology for Anomaly Detection in Buildings Energy Efficiency. Energy Reports 10: 244–54.

[4] Sosa Donoso, J. R., Flores, M., Naya, S., and Tarrío-Saavedra, J. (2023). Local Correlation Integral Approach for Anomaly Detection Using Functional Data. Mathematics, 11(4), 815.

[5] Carral, L., Tarrío-Saavedra, J., Sáenz, A. V., Bogle, J., Alemán, G., and Naya, S. (2021). Modelling operative and routine learning curves in manoeuvres in locks and in transit in the expanded Panama Canal. The Journal of Navigation, 74(3), 633–655.

[6] Barbeito, I., S. Zaragoza, J. Tarrío-Saavedra, and S. Naya. (2017). Assessing Thermal Comfort and Energy Efficiency in Buildings by Statistical Quality Control for Autocorrelated Data. Applied Energy 190: 1–17.

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

Desarrollo de Modelos Predictivos AutoML en Abanca

José Piñeiro Abal[1], Juan Manuel Mazaira Gómez[2]

RESUMO

En Abanca hemos desarrollado una herramienta interna de modelado predictivo basada en las herramientas AutoML. Estas herramientas consisten en la automatización de todo el ciclo de vida de construcción de modelos predictivos, desde la lectura del conjunto de datos de entrenamiento hasta la puesta en producción del mejor modelo. Esta herramienta está construida para tres tipos de problemas: clasificación binaria, clasificación multiclase y clasificación de texto.

Además, gracias al uso de RMarkdown y las librerías avanzadas de R, flexdashboard, knitr y plotly obtenemos un informe interactivo en formato .html que nos proporciona la parte de explicabilidad del modelo seleccionado.

Palabras y frases clave: AutoML, Modelos Predictivos.

1. Herramienta AutoML

En Abanca hemos desarrollado una amplia variedad de modelos predictivos que ayuden a detectar la necesidad o el interés de un cliente, desde la contratación de ciertos productos, el abandono de los mismos, etc. Para simplificar y acelerar el proceso de construcción de modelos, hemos creado una herramienta interna basada en las tecnologías de AutoML (Aprendizaje Automático Automatizado). Esta herramienta automatiza todo el ciclo de vida de desarrollo de modelos predictivos, desde la preparación de datos hasta la selección de algoritmos, la evaluación del rendimiento y selección del mejor modelo de manera automática. Está diseñada específicamente para abordar tres tipos de problemas: clasificación binaria, clasificación multiclase y clasificación de texto.
Los resultados que obtenemos con esta herramienta son los siguientes:

- **Eficiencia:** Reduce significativamente el tiempo y los recursos necesarios para desarrollar modelos predictivos, ya que automatiza la mayoría de los pasos del proceso.
- **Accesibilidad:** Permite que usuarios que no son expertos en aprendizaje automático utilicen modelos avanzados sin necesidad de conocimientos profundos en el campo.
- **Automatización completa:** Construcción de modelos sin necesidad de ejecutar código.
- **Flexibilidad:** Ofrece la capacidad de manejar diferentes tipos de variables objetivo y problemas de clasificación.
- **Mejora continua:** Proporciona una base sobre la cual se pueden realizar mejoras y ajustes adicionales.
- **Comunicación efectiva:** Permite la creación de informes interactivos y dinámicos que facilitan la comunicación de los resultados del modelo.
- **Puesta en producción automática:** Permite la puesta en producción del modelo seleccionado de manera automática.

El procedimiento consta de los siguientes pasos:

1. **Lectura de datos:** Realiza la lectura del tablón de entrenamiento de Teradata. Crea en caso de no existir la carpeta que le indicamos en el procedimiento y se tomará como base para los siguiente pasos. Tranforma los NAs según se le especificase en media, mediana o 0, en el caso de las variables factor crearía una categoría NA. Detecta variables categóricas y las transforma. Elimina caracteres atípicos. Crea un `.Rdata` con los datos del modelo.

2. **Estudio de las variables:** Lee el `.Rdata` creado en el paso anterior. Representación de las correlaciones de los datos. Representaciones univariantes de las variables de tipo numérico en la carpeta Numerico. Representaciones univariantes de las variables de tipo numérico en la carpeta Factor.

3. **Selección de variables:** Eliminación de variables correlacionadas, duplicadas y con muchos niveles y actualización del `.Rdata`.

4. **Entrenamiento de modelos ML:** Dependiendo de la variable objetivo permite realizar undersampling y oversampling mediante la eliminación de casos o mediante la técnica SMOTE. Entrenamiento utilizando las técnicas ML seleccionadas (RF, MLP, GBM, LightGBM y XGBoost) haciendo una búsqueda aleatoria o una aproximación bayesiana de hiperparámetros. Selección del mejor modelo evitando aquellos modelos que están sobreajustados. Creación del informe `.html` con el resumen del funcionamiento del modelo, variables más importantes, etc.

2. Generación del .html

Gracias al uso de librerías como `flexdashboard`, `knitr` y `plotly`, se puede ejecutar el procedimiento sin necesidad de ejecutar ningún tipo de código y además te permite generar un `.html` interactivo que ayude a explicar el procedimiento y el mejor modelo seleccionado.
Como se puede ver en la Figura 1 al darle a knitr tenemos la opción de Knitr with Parameters... en dónde podemos introducir el valor de las variables de manera manual cómo se puede ver en la Figura 2 y una vez que se modifiquen esos parámetros se ejecutaría el código finalizando con la construcción del html interactivo. Por ejemplo la Figura 3 es un ejemplo de un .html generado a través de estas librerías.

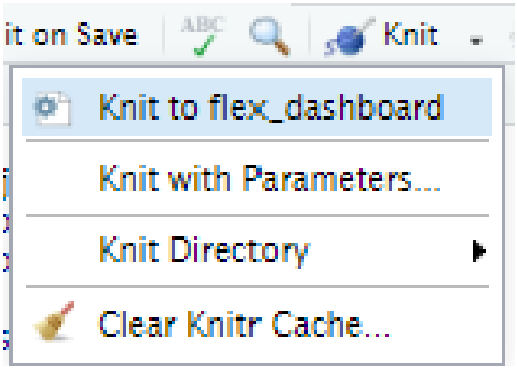


Figura 1: Ejecución de knitr con parámetros.

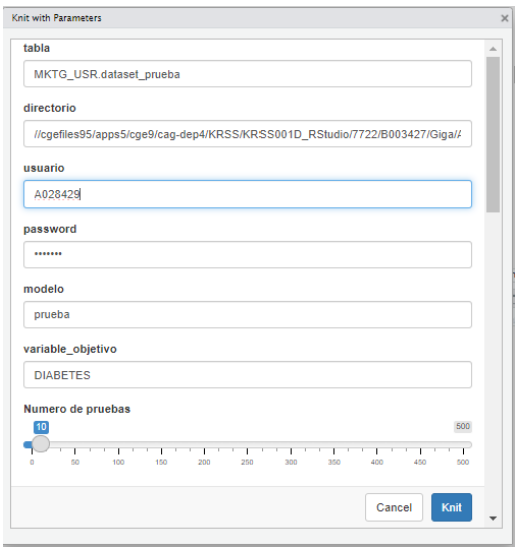


Figura 2: Introducción de parámetros.

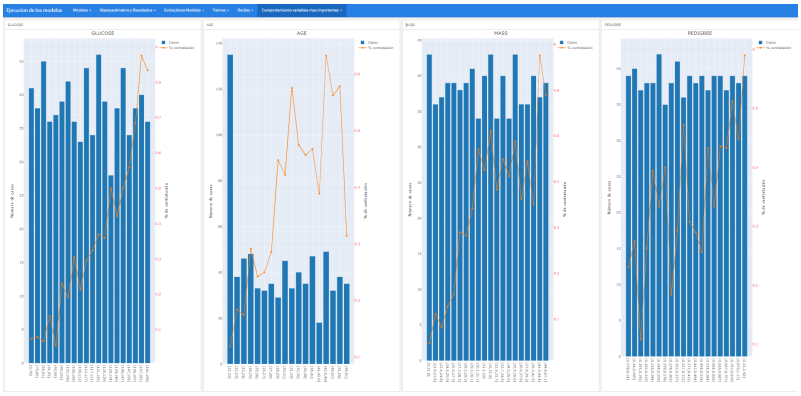


Figura 3: Ejemplo de html interactivo al finalizar el proceso.

Referencias

- Plotly. Sitio web oficial de Plotly: https://plotly.com/
- H2O. Documentación oficial de H2O: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html
- flexdashboard. Documentación oficial de flexdashboard: https://rmarkdown.rstudio.com/flexdashboard/
- XGBoost. Documentación oficial de XGBoost en R: https://xgboost.readthedocs.io/en/latest/R-package/index.html
- LightGBM. Documentación oficial de LightGBM en R: https://lightgbm.readthedocs.io/en/latest/R-API.html

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

LA (R)EVOLUCIÓN DE LOS ENTORNOS GRÁFICOS DE USUARIO (GUI) PARA R

Miguel Ángel Rodríguez Muiños[1], Teresa Seoane-Pillado[2]

[1] Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia
[2] Facultade de Ciencias da Saúde. Universidade de A Coruña.

RESUMO

R es una herramienta poderosa, pero, como cualquier lenguaje de programación, requiere de cierto esfuerzo en su aprendizaje. Muchos estadísticos, científicos de datos y/o programadores utilizan R directamente en la consola de comandos. Sin embargo, la línea de comandos puede resultar bastante desalentadora para un usuario medio o no-programador, incluso para usuarios que toman contacto por primera vez con R. Disponemos, desde hace años, de entornos gráficos de usario que nos facilitan el manejo del programa. Sin embargo, una de las asignaturas pendientes era poder ofrecer entornos de trabajo orientados, específicamente, al usuario no-programador. Afortunadamente, en la actualidad, disponemos de interfaces gráficas de usuario, de alta calidad, que ayudan a aplanar la curva de aprendizaje y falicitan la interacción con el programa y nos permiten utilizarlo sin tener que escribir las instrucciones en la linea de comandos. Haremos un repaso por los GUI, más destacados y modernos, especialmente diseñados para usar R como un software de alto nivel orientado análisis estadístico de datos. Entre ellos, encontramos soluciones como BlueSky Statistics (Figura 1) desarrollado por un equipo de antiguos miembros de SPSS; JAMOVI (Figura 2) desarrollado, como fork de JASP, por antiguos miembros de su equipo y JASP (Figura 3), de la Universidad de Amsterdam, entre otros.

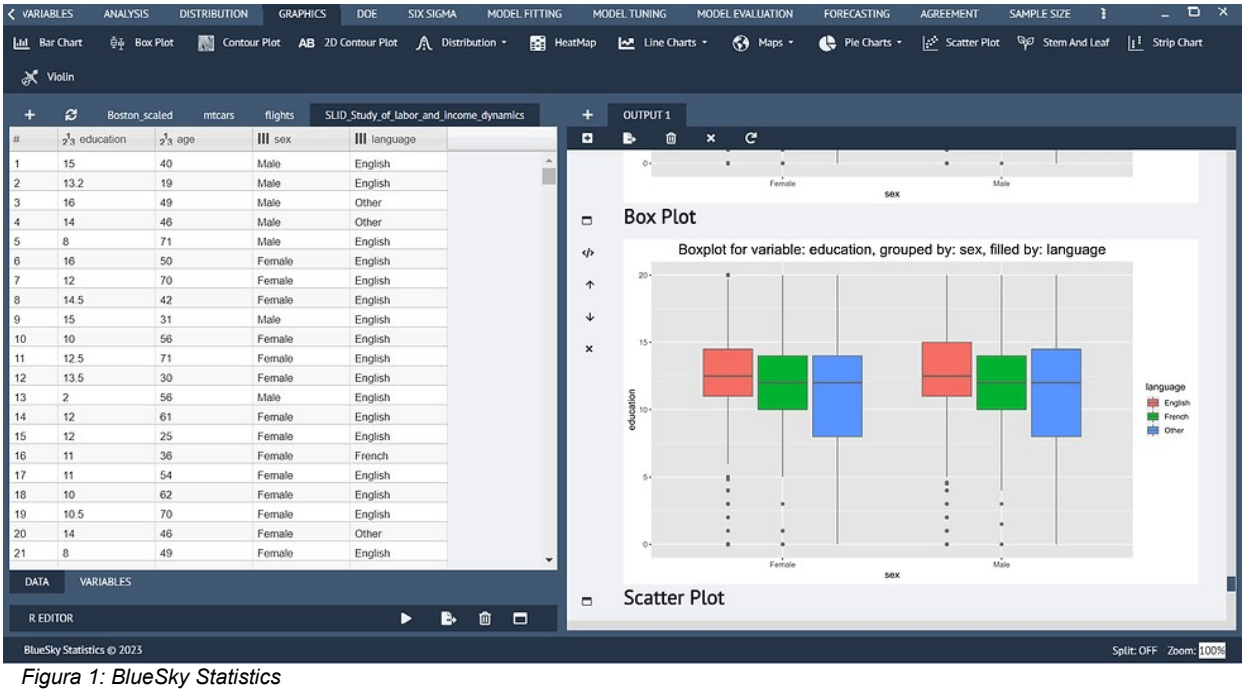


Figura 1: BlueSky Statistics

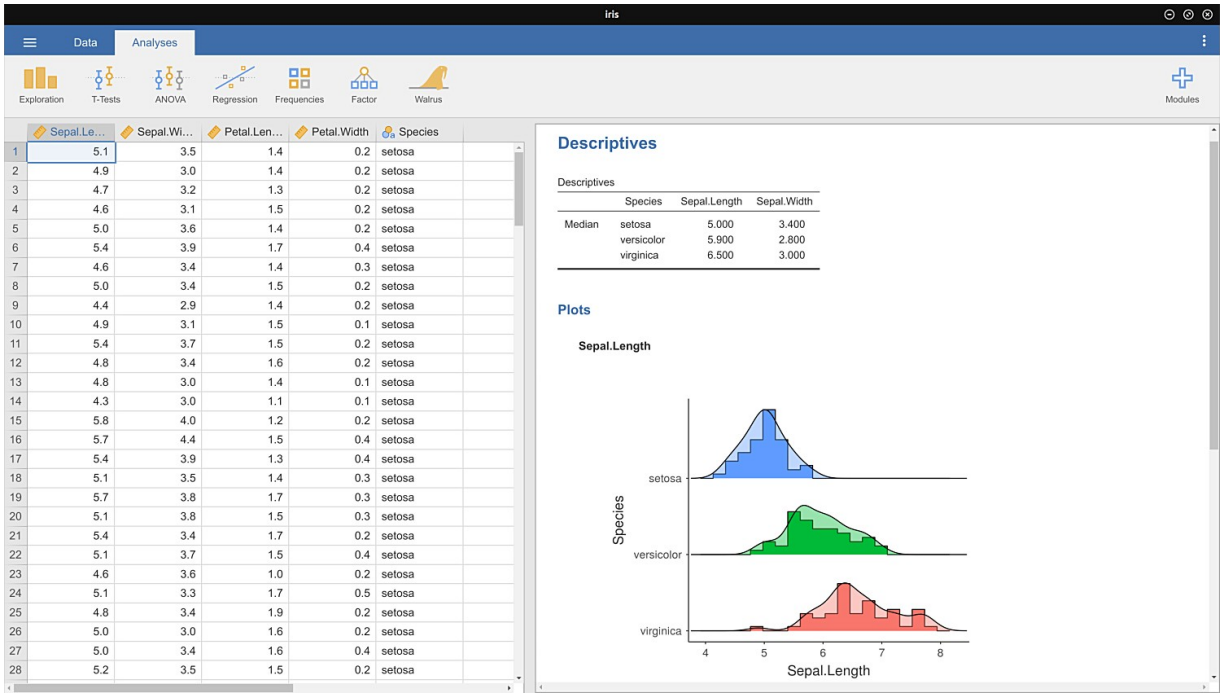


Figura 2: JAMOVI

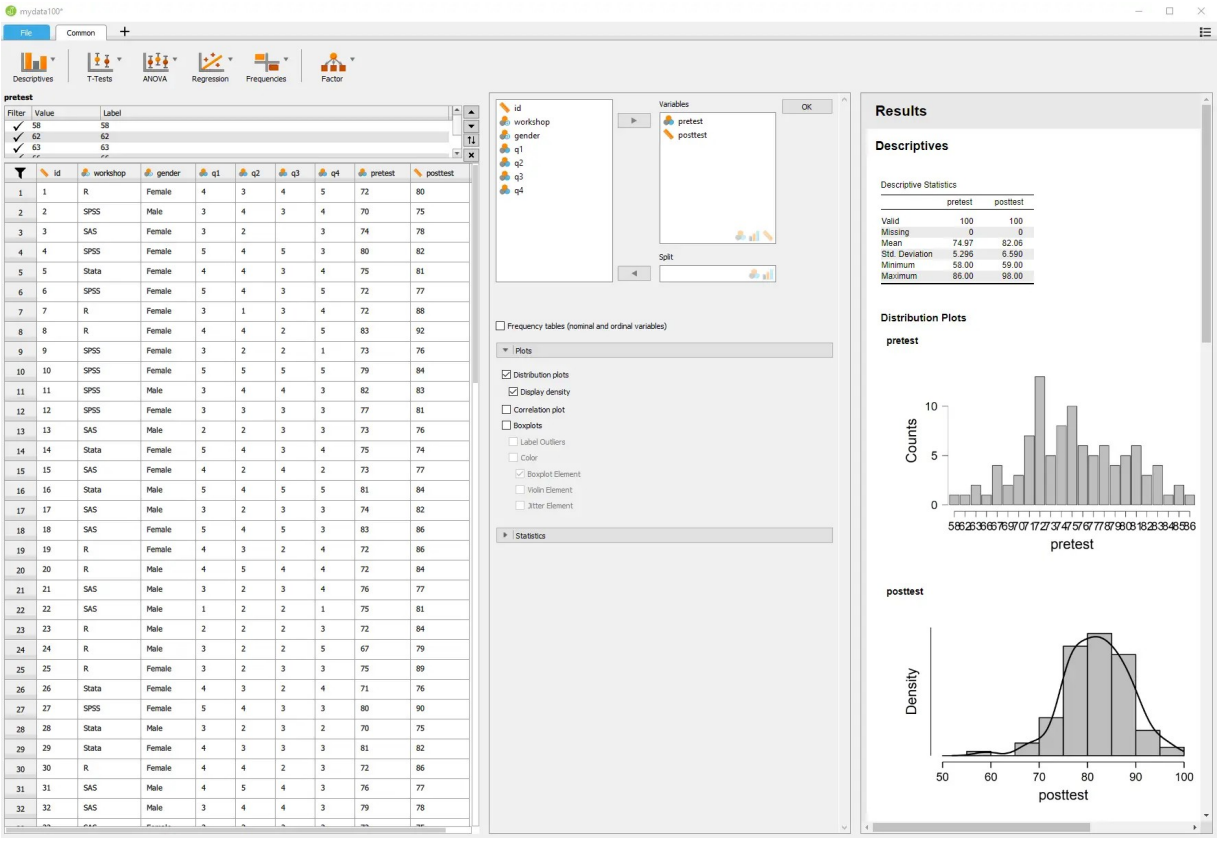


Figura 3: JASP

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

UNHA INTRODUCIÓN Ó PAQUETE `meteospain`

Marta Rodríguez Barreiro[1,2], María José Ginzo Villamayor[2,3]

[1]Universidade da Coruña, Departamento de Matemáticas
[2]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga)
[3]Universidade de Santiago de Compostela, Departamento de Estatística, Análise Matemática e Optimización

RESUMO

O paquete *meteospain* permite obter datos de diferentes servizos meteorolóxicos españois. A principal vantaxe deste paquete é que a forma de acceder ós servizos é uniforme e os datos que se obteñen son estandarizados e compatibles entre si. O paquete descárgase directamente desde o CRAN e instálase con facilidade. Neste traballo presentarase o paquete *meteospain*, as súas principais funcionalidades e amosarase un caso de uso no que se emprega o paquete no proceso de obter os vectores de desprazamento nun incendio forestal.

Palabras e frases chave: *meteospain*, datos meteorolóxicos, AEMET, MeteoGalicia, incendios forestais

1. INTRODUCIÓN

Os datos meteorolóxicos son fundamentais en proxectos ou estudios en diferentes ámbitos como a saúde pública, a xestión do risco, agricultura, pesca, xestión da auga, transporte, enerxía... Moitos servizos meteorolóxicos permiten obter os seus datos mediante APIs (interface de programación de aplicacións), pero a forma de descargar os datos e o formato no que se obteñen dependen de cada servizo.

A AEMET (Axencia Estatal de Meteoroloxía, [1]) é un organismo oficial de España que ten por obxecto a prestación de servizos meteorolóxicos. En total, hai arredor de 852 estacións automáticas de AEMET por toda España das que 57 están en Galicia. Ademais, conta cun portal de datos abertos que permite consultar parte da información elaborada por AEMET (`https://www.aemet.es/es/datos_abiertos`). Entre os datos que proporciona poden atoparse valores climatolóxicos normais no período 1981-2010, valores extremos absolutos desde o ano 1920, resumos sobre o estado do clima e a súa evolución, predicións por concellos a 7 días, predicións horarias por concellos (a 2 días vista)... Toda esta información pode atoparse no catálogo dos datos abertos de AEMET (`https://www.aemet.es/es/datos_abiertos/catalogo`). Porén, para determinados estudos, os datos proporcionados pola axencia poden non ser suficientes, ou non ter a precisión necesaria. Nestes casos, ás veces é necesario empregar máis dun portal meteorolóxico como fonte de datos. Por exemplo, no caso de Galicia, MeteoGalicia (a Unidade de Observación e Predición Meteorolóxica de Galicia, [5]) conta cun total de 169 estacións meteorolóxicas repartidas por todo o territorio galego. Combinar os datos obtidos de AEMET cos que proporciona MeteoGalicia pode ser unha boa solución. MeteoGalicia tamén proporciona acceso ós seus datos, pero a forma de acceder a eles para descargalos e o formato no que se obteñen é diferente ós da AEMET. Isto pode ser un problema ou pode complicar a utilización dos datos combinados de ambas fontes. O paquete *meteospain* ([3]) permite estandarizar isto, facilitando o emprego de datos meteorolóxicos de distintas fontes nun mesmo proxecto.

2. O PAQUETE `meteospain`

O paquete *meteospain* ([3]) permite acceder a diferentes estacións meteorolóxicas españolas dunha maneira uniformada, proporciona funcións que permiten obter os datos dos distintos portais meteorolóxicos tan só modificando algúns parámetros, e conserva o mesmo formato para todos os datos descargados.
O paquete atópase no CRAN e pode ser facilmente instalado mediante a sentencia:

```
install.packages('meteospain')
```

Ademais, o paquete conta cunha documentación moi elaborada e varios tutoriais (ou *vignettes*) nos que explican de xeito detallado como empregar as funcións do paquete, con exemplos.
O paquete conta con 5 funcións principais:

1. `get_meteo_from`: permite conectar e descargar datos dos distintos servizos meteorolóxicos.
2. `get_quota_from`: permite obter información da API utilizada no servizo meteorolóxico. Dependendo do servizo, algunhas APIs só permiten un número determinado de solicitudes de datos. Esta función accede ó número de solicitudes que se permiten no servizo para o usuario.
3. `get_stations_info_from`: permite obter información das estacións dos diferentes servizos.
4. `services_options`: mostra as opcións para acceder ós diferentes servizos meteorolóxicos.

Como se explicou anteriormente, todos os servizos poden consultarse a partir das mesmas funcións, o que permite estandarizar o seu uso.
Os servizos meteorolóxicos ós que se pode acceder desde *meteospain* son: AEMET ([1]), MeteoCat (servizo meteorolóxico de Cataluña, [7]), Meteoclimatic (rede non profesional de estacións meteorolóxicas automáticas repartidas por todo o territorio español, [4]), MeteoGalicia ([5]) e RIA (Red de Información Agroclimática de Andalucía, [6]).
A continuación mostrarase algúns exemplos dos datos que se poden obter dos diferentes servizos.

AEMET

AEMET ([1]) é o servizo nacional meteorolóxico que proporciona datos de calidade para uso público e estudos de investigación, así como produtos de predición e avisos de desastres. O paquete *meteospain* só accede á rede de estacións meteorolóxicas automáticas.
As resolucións temporais ás que se poden acceder son:

- `current_day`: devolve os datos das lecturas das últimas 24 horas das estacións seleccionadas.
- `daily`: devolve as medidas agregadas diarias de todas as estacións seleccionadas.
- `monthly`: devolve as medidas mensuais agregadas de unha estación no período indicado.
- `yearly`: devolve os valores anuais agregados de unha estación no período indicado.

Para acceder ós datos de AEMET é necesario ter unha chave de acceso persoal (*API key*). Para obter esta chave, é preciso acceder ó servizo *Open Data* de AEMET e seguir as instrucións (`https://opendata.aemet.es/centrodedescargas/inicio`).
meteospain non da opción á obtención desta chave directamente, pero pode facerse empregando outro paquete de R, *keyring* ([2]).

MeteoCat

MeteoCat ([7]) é o servizo meteorolóxico Catalán e ofrece acceso a datos meteorolóxicos. O paquete *meteospain* só accede á rede de estacións meteorolóxicas automáticas.
As resolucións temporais son:

- `instant`: proporciona as últimas 4 horas de medidas para as estacións seleccionadas.

- `hourly`: devolve todas as medidas (algunhas estacións en intervalos de 30 minutos, outras de 60, outras máis) para todas estacións seleccionadas.
- `daily`: devolve as medidas agregadas diarias para o mes proporcionado.
- `monthly`: devolve as medidas agregadas mensuais para o ano proporcionado.
- `yearly`: devolve os valores anuais agregados para todos os anos dispoñibles.

Do mesmo xeito que AEMET, MeteoCat precisa dunha chave persoal para poder acceder e descargar os datos. Para obter esta chave, pódense seguir as instrucións que se atopan en `https://apidocs.meteocat.gencat.cat/`. Tamén se pode utilizar o paquete *keyring* ([2]).

Meteoclimatic

Meteoclimatic ([4]) é unha rede non profesional de estacións automáticas repartidas por todo o territorio español. Non se garante a calidade dos datos, como no resto dos servizos.
No caso deste portal, só existen datos agregados do día actual, polo que non hai máis resolucións temporais. Ademais, non acepta múltiples estacións na mesma consulta.

MeteoGalicia

MeteoGalicia ([5]) ofrece datos da rede de estacións automáticas galegas.
Existen diferentes resolucións temporais:

- `instant`: devolve as últimas lecturas de datos das estacións seleccionadas.
- `current_day`: proporciona as últimas 24 horas de medidas para as estacións seleccionadas.
- `daily`: devolve as medidas agregadas diarias das estacións seleccionadas para o período de tempo indicado.
- `monthly`: devolve as medidas agregadas mensuais para o período de tempo indicado e as estacións seleccionadas.

RIA

A RIA ([6]) ofrece información da rede de estacións meteorolóxicas automáticas andaluzas.
Ofrece datos con diferentes resolucións temporais:

- `daily`: devolve as medidas agregadas diarias das estacións seleccionadas para o período de tempo indicado.
- `monthly`: devolve as medidas agregadas mensuais para o período de tempo indicado e as estacións seleccionadas.

3. CASO DE USO

No contexto dun proxecto de investigación dun incendio forestal no marco da Civil UAVs Inititative (CUI), desde CITMAga traballamos na obtención dos vectores de desprazamento dun incendio forestal (a tempo pasado), para, a posteriori, analizar o seu comportamento.
Para obter estes vectores de desprazamento, entre outras cousas, é preciso coñecer o vento en cada período de tempo que se quere analizar.
Neste contexto, empregouse o paquete *meteospain* para a obtención dos datos meteorolóxicos. Presentarase un pequeno exemplo de como foi empregado, e de como se pode empregar para futuras evolucións do algoritmo desenvolvido.

Referencias

[1] Agencia Estatal de Meteorología (AEMET), https://www.aemet.es/es/portada

[2] Csárdi, G. (2022). keyring: Access the System Credential Store from R. R package version 1.3.1, https://CRAN.R-project.org/package=keyring.

[3] Granda, V. (2023) Meteospain: Access to Spanish Meteorological Stations Services. R package version 0.1.2, https://CRAN.R-project.org/package=meteospain.

[4] Meteoclimatic, https://www.meteoclimatic.net/?screen_width=1280

[5] MeteoGalicia, https://www.meteogalicia.gal/web/inicio.action?request_locale=gl

[6] Red de Información Agroclimática de Andalucía (RIA), https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web/

[7] Servei Meteorlògic de Catalunya (Meteocat), https://www.meteo.cat/

X Xornada de Usuarios de R en Galicia
Santiago de Compostela, 18 de outubro do 2023

`IndexNumber`: USANDO R PARA MEDIR A EVOLUCIÓN DE MAGNITUDES

Alejandro Saavedra-Nieves[1] e Paula Saavedra-Nieves[1]

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

RESUMO

Os números índice son medidas estatísticas descritivas útiles no eido económico para comparar magnitudes simples e complexas rexistradas, polo xeral, en dos períodos de tempo. Aínda que estos conceptos teñen unha longa historia, seguen desempeñando un papel chave nas sociedades modernas actuais, nas que se dispón de grandes cantidades de datos económicos que están dispoñibles e precisan ser analizados. Neste relatorio presentaremos o paquete de R `IndexNumber`, con grandes capacidades para calcular aqueles números índice máis coñecidos na literatura.

Palabras e frases chave: Números índice, estatística descritiva, medida, ciencias sociais.

Referencias

[1] Saavedra-Nieves, A., Saavedra-Nieves, P. (2021). IndexNumber: An R Package for Measuring the Evolution of Magnitudes. *The R Journal* 13, 253-275.

[2] Saavedra-Nieves, A., Saavedra-Nieves, P. (2001). *IndexNumber: Index Numbers in Social Sciences*. R package version 1.3.2.

AUTORES

IX 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

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

ISBN 9 788409 551293