XI XORNADA DE USUARIOS DE R EN GALICIA

24 de outubro de 2024

LIBRO DE RESUMOS

> ORGANIZA

> PATRONCINAN

XI XORNADA DE USUARIOS DE R EN GALICIA

PROGRAMA E RESUMOS

24 de outubro de 2024

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

XI XORNADA DE USUARIOS DE R EN GALICIA

Editora: María José Ginzo Villamayor

ISBN: 978-84-09-66122-0

Presentación

A Asociación de usuarios de software libre da Terra de Melide (MeLiSA) comprácese en presentar a XI Xornada de Usuarios de R en Galicia.

Este evento busca converterse nun punto de encontro para todas aquelas persoas interesadas en compartir as súas experiencias e establecer colaboracións dentro da comunidade, ao tempo que promove e difunde o coñecemento libre da linguaxe estatística R e as súas aplicacións prácticas.

O programa contempla vinte relatorios ao longo de todo o día. Dos cales seis son convidados e ás outras catorce atenderon á chamada de recepción de propostas.

O evento contará con participantes destacados do Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga), da Xunta de Galicia, así como do Instituto Galego de Estatística e das tres universidades galegas ou da Universidad Carlos III de Madrid. Tamén participarán académicos de universidades internacionais, como a Universidad Cooperativa de Colombia, Universidade Federal Fluminense (Brasil) e da Academia da Força Aérea (Brasil), así como un profesor de Ensino Medio do IES Pedra da Auga en Ponteareas.

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

Santiago de Compostela, outubro de 2024

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

XI XORNADA DE USUARIOS DE R EN GALICIA

Data

24 de outubro de 2024

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 XI Xornada.

Programa

XI XORNADA DE USUARIOS DE R EN GALICIA

24 de outubro de 2024

09:10 – 09:30	Sesión de apertura Mª Elena Vázquez Cendón (Decana da Facultade de Matemáticas), Jose Ameijeiras Alonso (Coordinador do Máster en Técnicas Estatísticas – Universidade de Santiago de Compostela), Rafeael Rodríguez Gayoso (Tesoureiro de MeLiSA), Mª José Ginzo Villamayor (Presidenta de MeLiSA – Universidade de Santiago de Compostela),
09:30 – 09:50	*NPCirc*: novas actualizacións de métodos non paramétricos para datos circulares María Alonso Pena, Jose Ameijeiras Alonso, Rosa M. Crujeiras e Irène Gijbels. *Universidade de Santiago de Compostela*
09:50 – 10:10	R para a elaboración e visualización da estatística de vivendas familiares principais ocupados en Galicia R para a elaboración e visualización da estatística de vivendas familiares principais ocupados en Galicia. *Instituto Galego de Estatística*
10:10 – 10:30	Os efectos da pandemia na fenda dixital na educación en américa central e do sur José Manuel Amoedo, Bruno Blanco Varela e Hugo Campos Romero. *Universidade de Santiago de Compostela*
10:30 – 10:50	Efecto da asma na calidade de vida autopercibida polos pacientes en España Alba Paz Castro e José Manuel Amoedo. *Universidad de Santiago de Compostela*
10:50 – 11:10	R, Webscraping e Open Science. Esquivando balas en Matrix Álvaro Theotonio. *Universidad Carlos III de Madrid*
11:10 – 11:30	Comparison of cumulative incidence curves with multiple causes of death Nora Martínez Villanueva, Marta Sestelo, Luís Meira Machado e Javier Roca Padiñas. *Universidade de Vigo*
11:30 – 12:00	**PAUSA**
12:00 – 12:20	Keep the ball rolling: control estadístico de la calidad para la industria 5.0 Salvador Naya Fernández, Javier Tarrío Saavedra e Miguel Flores. *Universidade da Coruña*
12:20 – 12:40	Non metas a gamba! análises de datos xenómicos con R Adrián Casanova, Miguel Hermida, Paulino Martínez, Inmaculada Carrasco, Francisca Robles, Rafael Navajas, Roberto de la Herrán e Carmelo Ruiz. *Universidade de Santiago de Compostela*
12:40 – 13:00	*adegenet* e *parallelstructure*: análises xenómicos de estrutura poboacional en árnica Fernando Cabana, Adrián Casanova, Manuel A. Rodríguez Guitián, Andrés Blanco, Carlos Real, Rosa Romero, Carmen Bouza e Manuel Vera. *Universidade de Santiago de Compostela*
13:00 – 13:20	O papel de R en Investigación Mariña. Da xenómica a estudo global dos océanos Isabel Fuentes Santos. *Instituto de Investigacións Mariñas*
13:20 – 13:40	Calculo de rutas de escape nun incendio forestal Manuel Antonio Novo Pérez, Marta Rodríguez Barreiro, e María José Ginzo Villamayor. *Centro de Investigación e Tecnoloxía Matemática de Galicia*
13:40 – 14:00	Calculando a severidade dun incendio forestal con R Marta Rodríguez Barreiro, Manuel Antonio Novo Pérez, e María José Ginzo Villamayor. *Centro de Investigación e Tecnoloxía Matemática de Galicia*
14:00 – 16:20	**PAUSA**
16:20 – 16:40	El análisis por componentes en las ciencias sociales Jorge Alejandro Obando, Aura Viviana Rincón Ramirez, e Laura Nathalia Obando. *Universidad Cooperativa de Colombia (Colombia)*
16:40 – 17:00	Introduction to the R Packages based on JDemetra+ 3rd version Cheyenne Amoroso, Carolina García Martos, Germán Aneiros, José A. Vilar Fernández, Manuel Oviedo de la Fuente e Mario Francisco Fernández. *Universidade da Coruña*
17:00 – 17:20	Aplicativos para Aprendizagem Acelerada das Iteraçoes do Método Simplex Luciane Ferreira Alcoforado. *Academia da Força Aérea (Brasil)*
17:20 – 17:40	Toponomastics, uma ferramenta para o estudo da toponímia com foco no superestrato Afonso Xavier Canosa Rodrigues. *Xunta de Galicia*
17:40 – 18:00	Processo de otimização de uma carteira de ativos utilizando a linguagem R Ariel Levy, Marcus Antonio Cardoso Ramalho, e Eduardo Camilo da Silva. *Universidade Federal Fluminense (Brasil)*
18:00 – 18:10	**PAUSA**
18:10 – 18:30	Análise de datos contables mediante R. Exemplo de análise de certos datos de facturación e gastos na Xunta de Galicia Marcos Fernández Arias. *Xunta de Galicia*
18:30 – 18:50	Mandalas Matemáticas: Uma Releitura das Padronagens e Cores das Cerâmicas de Sargadelos Luciane Ferreira Alcoforado, João Paulo Martins dos Santos, Maria Cláudia e Jesús Machado. *Academia da Força Aérea (Brasil)*
18:50 – 19:10	Análisis de nubes de puntos masivas en R Nataly Romarís Lodeiro, Olamar Benavente Fernández, Rubén Fernández Casal e Salvador Naya Fernández. *Universidade da Coruña*
19:30 – 19:35	Clausura María José Ginzo Villamayor. *Universidade de Santiago de Compostela*

Índice

npcirc: NOVAS ACTUALIZACIÓNS DE MÉTODOS NON PARAMÉTRICOS PARA DATOS CIRCULARES. María Alonso Pena, Jose Ameijeiras Alonso, Rosa M. Crujeiras e Irène Gijbels. Universidade de Santiago de Compostela 11

R PARA A ELABORACIÓN E VISUALIZACIÓN DA ESTATÍSTICA DE VIVENDAS FAMILIARES PRINCIPAIS OCUPADOS EN GALICIA. Esther López Vizcaíno, Isabel del Río Viqueira e Solmary Silveira Calviño. Instituto Galego de Estatística 46

OS EFECTOS DA PANDEMIA NA FENDA DIXITAL NA EDUCACIÓN EN AMÉRICA CENTRAL E DO SUR. José Manuel Amoedo, Bruno Blanco Varela e Hugo Campos-Romero. Universidade de Santiago de Compostela..... 13

EFECTO DA ASMA NA CALIDADE DE VIDA AUTOPERCIBIDA POLOS PACIENTES EN ESPAÑA. Alba Paz Castro e José Manuel Amoedo. Universidade de Santiago de Compostela..... 67

R, WEBSCRAPING E OPEN SCIENCE. ESQUIVANDO BALAS EN MATRIX. Álvaro Theotonio. Universidad Carlos III de Madrid 74

COMPARISON OF CUMULATIVE INCIDENCE CURVES WITH MULTIPLE CAUSES OF DEATH. Nora Martínez Villanueva, Marta Sestelo, Luís Meira Machado e Javier Roca Padiñas. Universidade de Vigo..... 50

KEEP THE BALL ROLLING: CONTROL ESTADÍSTICO DE LA CALIDAD PARA LA INDUSTRIA 5.0. Salvador Naya Fernández, Javier Tarrío Saavedra e Miguel Flores. Universidade da Coruña 55

NON METAS A GAMBA! ANÁLISES DE DATOS XENÓMICOS CON R. Adrián Casanova, Miguel Hermida, Paulino Martínez, Inmaculada Carrasco, Francisca Robles, Rafael Navajas, Roberto de la Herrán e Carmelo Ruiz. Universidade de Santiago de Compostela 23

adegenet E parallelstructure: ANÁLISES XENÓMICOS DE ESTRUTURA POBOACIONAL EN ÁRNICA. Fernando Cabana, Adrián Casanova, Manuel A. Rodríguez-Guitian, Andrés Blanco, Carlos Real, Rosa Romero, Carmen Bouza e Manuel Vera. Universidade de Santiago de Compostela 19

O PAPEL DE R EN INVESTIGACIÓN MARIÑA. DA XENÓNMICA A ESTUDO GLOBAL DOS OCÉANOS. Isabel Fuentes Santos. Instituto de Investigacións Mariñas 41

CALCULO DE RUTAS DE ESCAPE NUN INCENDIO FORESTAL. Manuel Antonio Novo Pérez, Marta Rodríguez Barreiro e María José Ginzo Villamayor. Centro de Investigación e Tecnoloxía Matemática de Galicia 59

CALCULANDO A SEVERIDADE DUN INCENDIO FORESTAL CON R. Marta Rodríguez Barreiro, Manuel Antonio Novo Pérez e María José Ginzo Villamayor. Centro de Investigación e Tecnoloxía Matemática de Galicia 68

EL ANÁLISIS POR COMPONENTES EN LAS CIENCIAS SOCIALES. Jorge Alejandro Obando, Aura Viviana Rincón Ramirez e Laura Nathalia Obando. Universidad Cooperativa de Colombia..... 63

INTRODUCTION TO THE R PACKAGES BASED ON JDEMETRA+ 3RD VERSION. Cheyenne Amoroso, Carolina García Martos, Germán Aneiros, José A. Vilar Fernández, Manuel Oviedo de la Fuente, Mario Francisco Fernández. Universidade da Coruña 17

APLICATIVOS PARA APRENDIZAGEM ACELERADA DAS ITERAÇOES DO MÉTODO SIMPLEX. Luciane Ferreira Alcoforado. Academia da Força Aérea,..... 37

TOPONOMASTICS, UMA FERRAMENTA PARA O ESTUDO DA TOPONÍMIA COM FOCO NO SUPERESTRATO. Afonso Xavier Canosa Rodrigues. Xunta de Galicia 23

PROCESSO DE OTIMIZAÇÃO DE UMA CARTEIRA DE ATIVOS UTILIZANDO A LINGUAGEM R. Ariel Levy, Marcus Antonio Cardoso Ramalho, e Eduardo Camilo da SIlva. Universidade Federal Fluminense 42

ANÁLISE DE DATOS CONTABLES MEDIANTE R. EXEMPLO DE ANÁLISE DE CERTOS DATOS DE FACTURACIÓN E GASTOS NA XUNTA DE GALICIA. Marcos Fernández Arias. Xunta de Galicia .. 31

MANDALAS MATEMÁTICAS: UMA RELEITURA DAS PADRONAGENS E CORES DAS CERÂMICAS DE SARGADELOS. Luciane Ferreira Alcoforado, João Paulo Martins dos Santos, e Maria Cláudia de Jesus Machado. Academia da Força Aérea, ... 51

ANÁLISIS DE NUBES DE PUNTOS MASIVAS EN R. Nataly Romarís Lodeiro, Olamar Benavente Fernández, Rubén Fernández Casal e Salvador Naya Fernández. Universidade da Coruña..... 71

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

NPCirc: novas actualizacións de métodos non paramétricos para datos circulares

María Alonso Pena[1], Jose Ameijeiras Alonso[2], Rosa M. Crujeiras[2] e Irène Gijbels[3]

[1]Universidade de Santiago de Compostela
[2]CITMAga, Universidade de Santiago de Compostela
[3]Department of Mathematics, KU Leuven

RESUMO

O paquete `NPCirc`, inicialmente publicado no ano 2013, é un proxecto colaborativo no que se proporcionan ferramentas para realizar inferencia non paramétrica con datos circulares. O paquete contén funcións relacionadas con problemas de estimación da densidade e da función de regresión, cando algunhas das variables involucradas é de natureza circular. Nos últimos anos, engadíronse técnicas avanzadas como novos métodos de selección do parámetro de suavizado, técnicas de regresión avanzada e contrastes sobre a función de regresión.

Palabras e frases chave: Datos circulares, Densidade non paramétrica Regresión non paramétrica, Suavizado, Técnicas tipo núcleo.

1. INTRODUCIÓN

Os datos circulares son aqueles que teñen como soporte a circunferencia unidade: direccións, ángulos ou observacións periódicas. Este tipo de datos non poden ser analizados con técnicas estatísticas clásicas, debido á súa periodicidade. O paquete de `R` `NPCirc`, publicado inicialmente no ano 2013 (Oliveira et al, 2014), contén esencialmente métodos específicos para datos circulares, que se poden clasificar en tres grandes bloques:

- Conxuntos de datos reais contendo variables circulares.
- Funcións para estimar, non parametricamente, a densidade dunha variable circular.
- Funcións para estimar, de xeito non paramétrico, a función de regresión cando algunha das variables é circular.

Nos últimos anos, o paquete `NPCirc` ten sufrido distintas actualizacións, nos que se teñen engadido importantes funcións correspondentes aos distintos bloques.

2. NOVAS METODOLOXÍAS

As versións iniciais de `NPCirc` incluían a función `kern.den.circ`, que permitía obter o estimador non paramétrico da función de densidade circular. Xunto con esta, achegábanse funcións para seleccionar o parámetro de suavizado como `bw.pi, bw.rt, bw.CV, bw.boot`. Na última actualización, incorporouse a función `bw.AA`, implementando o método de Ameijeiras-Alonso (2024) baseado nas ideas de Sheather e Jones.
Por outra banda, as primeiras versións do paquete incluían as funcións `kern.reg.circ.lin`, `kern.reg.lin.circ` e `kern.reg.circ.circ` para estimar a función de regresión. Dentro das novas actualizacións, atópanse a incorporación de ferramentas para testar se, nestes contextos de regresión, existe un efecto significativo da variable explicativa sobre a resposta, mediante

`noeffect.circ.lin`, `noeffect.lin.circ` e `noeffect.circ.circ`. Ademais, a comparación de función de regresión en distintos grupos pódese levar a cabo mediante contrastes de hipóteses con `ancova.circ.lin`, `ancova.lin.circ` e `ancova.circ.circ` (Alonso-Pena et al, 2021).
Outros tipos de regresión máis xeralizados tamén son posibles grazas á función `circ.local.lik`, onde se asume unha variable explicativa circular e unha resposta que pode seguir unha resposta Bernoulli, Poisson ou gamma (Alonso-Pena et al, 2023).
Por último, dentro das técnicas de regresión, engadíronse ferramentas que non estiman a media condicional, senón as modas condicionais (Alonso-Pena e Crujeiras, 2023). Isto pódese levar a cabo mediante as funcións `modalreg.circ.lin`, `modalreg.lin.circ` e `modalreg.circ.circ`.

3. CONCLUSIÓNS

O paquete `NPCirc` considérase como un proxecto amplo onde se pretende proporcionar, á comunidade científica, ferramentas para a análise non paramétrica de datos circulares. Ademais das últimas actualizacións, espérase que no futuro próximo se engadan máis ferramentas como clustering modal non paramétrico, contrastes de bondade de axuste, tanto para densidade como para regresión, ou regresión tipo single-index.

AGRADECEMENTOS

Os autores agradecen a María Oliveira a cesión do mantemento do paquete. Ademais, agradécese o financiamento dos proxectos PID2020-116587GBI00, financiado por MCIN/AEI/10.13039/501100011033; ED431C 2021/24 da Xunta de Galicia e C16/20/002 da Research Fund KU Leuven, Belgium.

Referencias

[1] Alonso-Pena, M., Ameijeiras-Alonso, J. and Crujeiras, R.M. (2021). Nonparametric tests for circular regression. *Journal of Statistical Computation and Simulation* 91, 477–500.

[2] Alonso-Pena, M. and Crujeiras, R.M. (2023). Analyzing animal escape data with circular nonparametric multimodal regression. *Annals of Applied Statistics* 17 130–152.

[3] Alonso-Pena, M., Gijbels, I. and Crujeiras, R.M. (2023). A general framework for circular local likelihood regression. *Journal of the American Statistical Association.*

[4] Ameijeiras-Alonso, J. (2024). A reliable data-based smoothing parameter selection method for circular kernel estimation. *Statistics and Computing* 34.

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

ESTUDO DOS EFECTOS DA PANDEMIA DA COVID-19 NA FENDA DIXITAL EDUCATIVA EMPREGANDO A LIBRERÍA MATCHIT

José Manuel Amoedo[1], Bruno Blanco-Varela[2] e Hugo Campos-Romero[2]

[1] Grupo de Investigación ICEDE, Departamento de Economía Aplicada, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela, Santiago de Compostela, España

[2] Grupo de Investigación ICEDE, Departamento de Economía Cuantitativa, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela, Santiago de Compostela, España

RESUMO

Os sistemas educativos de todo o mundo víronse fortemente afectados pola Pandemia de Covid-19 e as políticas levadas a cabo para evitar a expansión da pandemia. Porén, estes impactos non parecen ter sido homoxéneos debido, entre outras cuestións, ó importante papel das Tecnoloxías da Información e a Comunicación (TIC) e ó diferente acceso a elas segundo o nivel socioeconómico dos fogares. O cal cobra aínda máis relevancia nos países emerxentes ou en vías de desenvolvemento. Nesta investigación estudamos o caso de oito países de América central e do sur co obxectivo de comprobar se a fenda no desempeño académico xerada polo diferente acceso ás TIC incrementouse durante a pandemia. Para isto empregamos datos da base PISA e diferentes librerías de R tales como dplyr, MathcIt, stats e ggplot2. Os resultados amosan como na meirande parte dos países estudados a pandemia incrementou considerablemente a fenda dixital.

Palabras e frases chave: tecnoloxías da información e a comunicación, fenda dixital, sistema educativo, América central e do sur, Covid-19

1. INTRODUCIÓN

Nos últimos anos, a integración das Tecnoloxías da Información e a Comunicación (TIC) transformou profundamente a educación, destacando a importancia de desenvolver competencias dixitais tanto para o éxito académico como para as oportunidades laborais. Porén, aínda que as institucións educativas comezaron a incorporar recursos dixitais no seus procesos de ensinanza, a dispoñibilidade destes recursos non garante o seu uso efectivo nin o desenvolvemento axeitado das competencias dixitais nin que todos os estudantes teñan un acceso similar a estes recursos.

A Pandemia da Covid-19 acelerou a dixitalización na educación, pero tamén puxo en evidencia e, en moitos casos, agravou as desigualdades preexistentes. A educación a distancia converteuse nunha solución necesaria, pero puido ter ampliado as fendas sociais no ámbito educativo [1]. Entre o anos 2018 e 2022, moitos países experimentaron un descenso no desempeño académico en competencias como matemáticas e lectura, o cal debe entenderse no contexto do impacto global da pandemia nos

sistemas educativos. Ademais, as respostas a esta crise e os seus efectos na equidade educativa variaron significativamente entre rexións e países.

Nesta investigación levamos a cabo un estudo para oito países de América central e do sur (Arxentina, Brasil, Chile, Colombia, República Dominicana, Panamá, Perú e Uruguai) de cara a comprobar se a fenda no rendemento entre estudantes xerada pola fenda dixital incrementouse nestes países polos efectos da pandemia. Para isto empregamos datos de dúas edicións do Programa para a Avaliación Internacional dos Estudantes (PISA) para o último ano dispoñible antes da Pandemia de Covid-19 [2] e para o primeiro dispoñible tras ela [3]. De cara a eliminar os problemas de endoxeneidade asociados ó diferente acceso ás TIC empregamos Propensity Score Matching (PSM), unha metodoloxía empregada frecuentemente neste tipo de estudos [4,5]. Para levar a cabo esta análise empréganse diferentes librerías de R tales como dplyr e MarchIt para construír a base de datos e levar a cabo o emparellamento.

Esta investigación divídese en catro seccións. Incluíndo esta introdución. Na segunda sección presentamos as variables e metodoloxía empregadas. A continuación, presentamos os resultados obtidos sobre a fenda entre estudantes con e sen acceso ás TIC para as competencias de ciencia, matemáticas e lectura. Na cuarta sección presentamos as principais conclusións obtidas desta investigación e as implicacións para as políticas públicas.

2. DATOS E METODOLOXÍA

Partindo das bases de datos comprendidas nas dúas edicións de PISA podemos conformar as covariables de tratamento, control e resultados recollidas na Táboa 1. As cales son empregadas para obter grupos de tratamento e control similares e, a partir das súas diferenzas, estimar os efectos do acceso ás TIC no desempeño académico. A covariable de tratamento empregada (PC_INTERNET) recolle se o estudante ten acceso a un ordenador para as tarefas escolares no fogar e acceso a internet (non inclúe a do teléfono móbil).

Grupo	Variable	Código
Tratamento	Ordenador dispoñible no fogar para facer os deberes	PC
	Acceso a internet (sen incluír a do teléfono móbil	INTERNET
	Ordenador para deberes e acceso a internet (PC*INTERNET)	PC_INTERNET
Control	Idade do estudante	AGE
	Sexo do estudante	SEX
	Estudante que repetiu algún curso	REPEAT
	Idade á que empezou os estudos oficiais	STU_BEGIN
	Orixe do estudante (inmigrante de 1º xeración, 2º xeración ou nativo	ORIGIN
	ISEC do estudante	ESCS
	Habitación propia	ROOM
	Tipo de centro segundo o réxime de propiedade	SCHLTYPE
	Tamaño do municipio ó que pertence o centro (en habitantes)	SCHLCOMSIZE
	Tamaño do centro (en estudantes)	SCHSIZE
	Tamaño medio das clases (estudantes na clase de lingua)	CLSIZE
	Alumnos por profesor	STRATIO
	Uso do agrupamento escolar no centro	ABGROUPING
Resultados	Media dos valores plausibles obtidos en ciencia	SCIENCE
	Media dos valores plausibles obtidos en lectura	READING
	Media dos valores plausibles obtidos en matemáticas	MATHEMATICS

Nota: o ISEC é un índice sintético recollido por PISA que mide o nivel social, económico e cultural do estudante
Táboa 1: variables de tratamento, control e resultados empregadas no estudo a partir da información dispoñible en PISA [2,3]

Para obter grupos de tratamento e control similares empregamos PSM e, máis concretamente, a metodoloxía do Veciño Máis Próximo (VMP) en combinación co

emparellamento exacto para algunhas variables clave (SEX, REPEAT, ORIGIN and SCHLTYPE). Para obter os mellores emparellamentos posibles optamos por empregar diferentes alternativas do VMP 1:k (con k=1, 3, 5 e 10) e seleccionamos para cada país e cada ano o emparellamento coas mellores medidas de equilibrio (nesgos estandarizados, pseudo-R^2 e análise gráfica). Para estimar a distancia empregamos o Modelo Lineal Xeneralizado (MLX).

Para levar a cabo esta análise primeiro constrúese a base de datos empregando a librería dplyr e conformar as covariables de tratamento, control e resultados. A continuación, emprégase a librería MatchIt para realizar o emparellamento para cada país e obter as medidas de equilibrio. En terceiro lugar, emprégase a librería stats para levar a cabo a análise estatística. Finalmente, de cara a presentar os resultados de forma gráfica utilízase a librería ggplot2.

Unha vez realizados os emparellamentos, a fenda no desempeño académico xerada pola fenda dixital é estimada a partir da diferencia entre as medias ponderadas para o grupo de tratamento e o grupo de control. O que é posible dado que ó eliminar os problemas de endoxeneidade mediante o emparellamento a única diferenza entre os estudantes é o feito de ter acceso ou non a un ordenador e internet (PC_INTERNET).

3. RESULTADOS

Os resultados obtidos recóllense na Figuras 1, que recolle a fenda no desempeño académico nos anos 2018 e 2022 nas competencias de ciencia, lectura e matemáticas. Nelas podemos observar como na meirande parte dos países a Pandemia de Covid-19 incrementou a fenda no desempeño entre os estudantes con acceso a ordenador e internet e aqueles sen acceso a ambos.

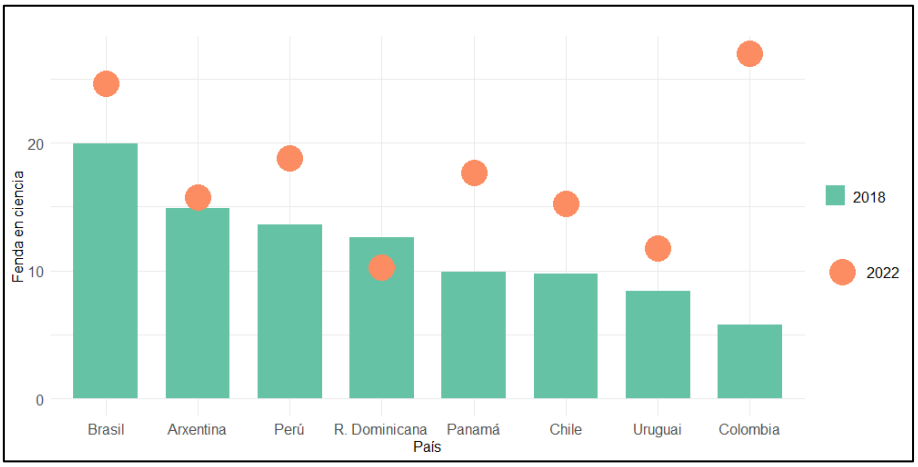


Figura 1: Fenda en ciencia

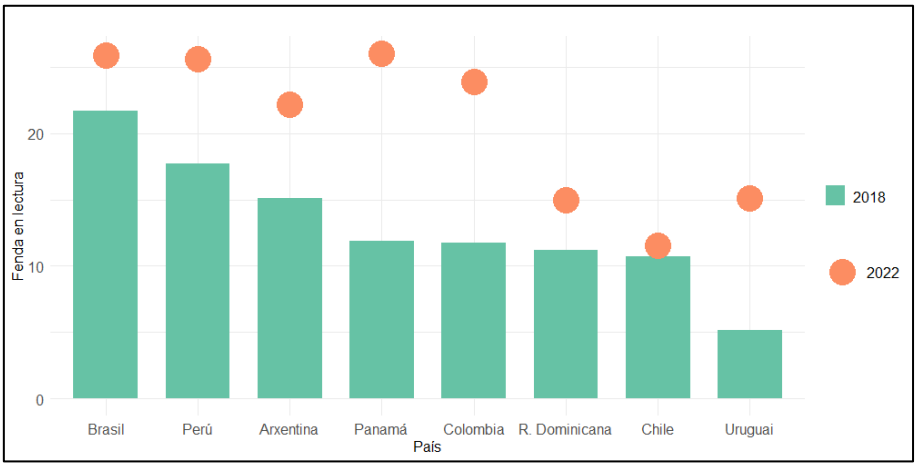


Figura 2: Fenda en lectura

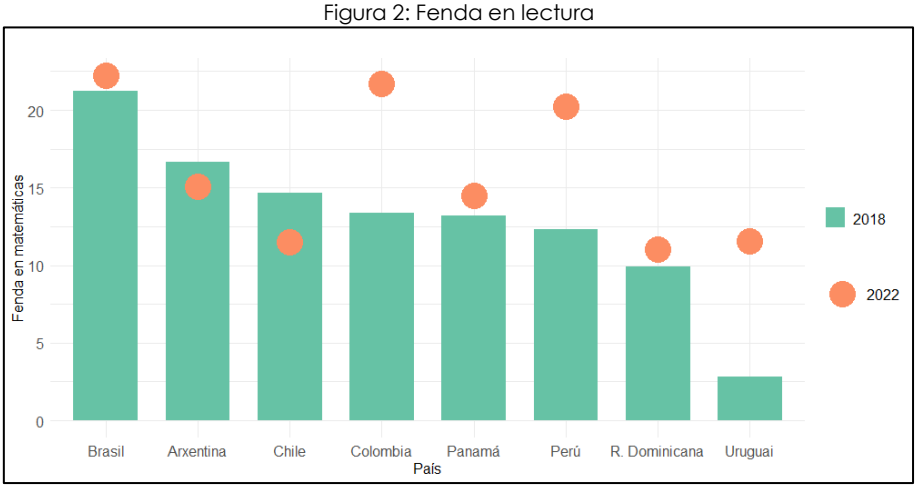


Figura 3: Fenda en matemáticas

4. CONCLUSIÓNS

A principal conclusión desta investigación é que a Pandemia de Covid-19 incremento a fenda no desempeño académico entre os estudantes con acceso a ordenador e conexión a internet na meirande nos países analizados. Se ben é preciso sinalar que existen certos matices e diferenzas entre os países e as competencias analizadas xa que os incrementos observados son dispares entre países e entre competencias. De feito, podemos observar como nalgúns países e competencias non existe un incremento significativo en dita fenda.

As implicacións derivadas deste estudo para as políticas públicas son a necesidade de fomentar o igual acceso ás TIC entre os estudantes para garantir a equidade no sistema educativo e a relevancia de ter en conta a desigual exposición ante fenómenos adversos, como o foi a Pandemia da Covid-19, dos estudantes segundo o seu contexto de cara a protexer a aqueles máis vulnerables.

Referencias

[1] Ogundari, K. (2023). Student access to technology at home and learning hours during COVID-19 in the U.S. Educational Research for Policy and Practice, 22(3), 443–460.

[2] OECD (2019). PISA 2018 Database. https://www.oecd.org/pisa/data/2018database/

[3] OECD (2023). PISA 2022 Database. https://www.oecd.org/pisa/data/2022database/

[4] Blanco-Varela, B., Amoedo, J. M., & Sánchez-Carreira, M. C. (2024). Analysing ability grouping in secondary school: A way to improve academic performance and mitigate educational inequalities in Spain?. International Journal of Educational Development, 107, 103028.

[5] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science : A Review Journal of the Institute of Mathematical Statistics, 25(1), 1–21.

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Introduction to the R Packages based on JDemetra+ 3rd version

Cheyenne Amoroso[1], Carolina García-Martos[2], Germán Aneiros[1], José A. Vilar[1], Manuel Oviedo de la Fuente[1], Mario Francisco-Fernández[1]

[1] CITIC, Grupo MODES, Departamento de Matemáticas, Universidade da Coruña.
[2] Escuela Tecnica Superior de Ingenieros Industriales (ETSII), Universidad Politécnica de Madrid.

ABSTRACT

Time series can exhibit a variety of underlying patterns that contribute to the observed changes over a period of time. These patterns are crucial to understanding the dynamics within the data and can be broken down into three main components: trend, seasonal and irregular. The trend component represents the long-term movement in the data, often indicating growth or decline over an extended period. The seasonal component captures periodic fluctuations that occur at regular intervals, such as monthly or quarterly, reflecting recurring events such as holidays or seasonal weather patterns. Finally, the irregular component accounts for the random or unpredictable variations in the data that cannot be attributed to the trend or seasonal components.

A common task in economics is the seasonal adjustment of time series, which involves removing the seasonal component from the data. This adjustment is crucial because seasonal fluctuations can obscure both short-term and long-term movements in the data, making it difficult to identify underlying trends and patterns. By removing the seasonal component, analysts can gain a clearer understanding of the underlying phenomena driving the series, allowing for more accurate forecasting and policy analysis.

In 2009, the European Statistical System (ESS) published Guidelines on Seasonal Adjustment [1] with the aim of harmonising European practices and improving the comparability of national infra-annual statistics. Following the first edition in 2009, the revised ESS Guidelines on Seasonal Adjustment were published in 2015 [2], presenting theoretical aspects and practical implementation issues in a user-friendly and easy-to-read framework. In line with the Eurostat guidelines, the National Statistics Institute (Spain) has established certain recommendations for the treatment of socio-economic time series [4].

The two most commonly used approaches to seasonal adjustment are ARIMA model-based adjustment and fixed filter-based adjustment. Both methods are recognised as equally valid by the ESS. Currently, INE applies seasonal adjustment using the ARIMA model-based approach. Specifically, the seasonal adjustment of time series is carried out using the Tramo-Seats method [3], which is a well-established procedure in the field of time series analysis and consists of two main stages. In the first step, a Reg-ARIMA model is fitted to the data to remove outliers and to account for calendar effects, such as the influence of holidays or different month lengths. This step is critical to isolate the deterministic effects that can bias the analysis. In the second stage, the signal is extracted using Wiener-Kolmogorov filtering, a sophisticated technique for separating the signal from the noise.

The recommended seasonal adjustment methods are currently implemented in several ways. In particular, INE recommends the use of the JDmetra+ package, which is also

widely recommended by Eurostat. JDemetra+ is an open source software for seasonal adjustment and time series analysis, developed by the National Bank of Belgium in the framework of Eurostat's "Centre of Excellence on Statistical Methods and Tools" with the support of the Deutsche Bundesbank and Insee.

There are several versions, including version 2.2.4, which is highly recommended, and the version 3.x family, which includes advanced features for seasonal adjustment and trend estimation, including high frequency data. The latest version 3.2.4 was released on 11 July 2024. The core Java algorithms of JDemetra+ can be accessed from within R. During the *XI Xornada de Usuarios de R* congress an overview of the R ecosystem in the context of JDemetra+ 3.x [5] is provided.

Keywords: Seasonal adjustment, Tramo-Seats, JDemetra+.

Acknowledgements: CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

References

[1] Eurostat (2009). *ESS Guidelines on Seasonal Adjustment: 2009 edition.*

[2] Eurostat (2015). *ESS Guidelines on Seasonal Adjustment: 2015 edition.*

[3] Gómez, Victor and Maravall, Agustin (1996). *Programs tramo and seats, instruction for user (beta version: september 1996).* Working papers, Banco de España.

[4] INE (2019). *Estándar del INE para la corrección de efectos estacionales y efectos de calendario en las series coyunturales.*

[5] Palate, Jean and Alain Quartier la Tente (2024). *rjd3tramoseats: Seasonal Adjustment with TRAMO-SEATS and 'JDemetra+ 3.0'*, R package version 3.2.2. `https://github.com/rjdemetra/rjd3tramoseats`

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

ADEGENET E PARALLELSTRUCTURE: ANÁLISES XENÓMICOS DE ESTRUTURA POBOACIONAL EN ÁRNICA

Fernando Cabana[1], Adrián Casanova[1], Manuel A. Rodríguez-Guitian[2], Andrés Blanco[1], Carlos Real[3], Rosa Romero[2], Carmen Bouza[1], Manuel Vera[1]

[1] Departamento de Zooloxía, Xenética e Antropoloxía Física, Facultade de Veterinaria, Universidade de Santiago de Compostela. Campus Terra, 27002 Lugo, España
[2] Departamento de Produción Vexetal, Escola Politécnica Superior, Universidade de Santiago de Compostela. Campus Terra, 27002 Lugo, España
[3] Departamento de Ecoloxía, Escola Politécnica Superior, Universidade de Santiago de Compostela. Campus Terra, 27002 Lugo, España

RESUMO

Dentro do amplo catálogo de paquetes do entorno R, moitos son fundamentais para a realización de análises xenómicas. Un exemplo destas análises serían aquelas que procuran definir as estruturas poboacionais dun conxunto de mostras, para a xestión dos recursos naturais. Neste caso de estudo móstrase o uso de dous paquetes de R, adegenet e ParallelStructure, para determinar a estrutura xenómica poboacional de 12 localidades de *Arnica montana*, planta medicinal con propiedades antiinflamatorias e de grande interese de conservación en Galicia e Europa. Estes paquetes ofrecen dúas aproximacións diferentes: (i) adegenet, baseada en análises discriminante de compoñentes principais e (ii) ParallelStructure o cal optimiza o programa de análise bayesiano STRUCTURE para poder procesar un gran volume de datos xenómicos.

Palabras e frases chave: *Arnica montana*, xenómica, estrutura poboacional.

1. INTRODUCIÓN

Arnica montana L. é unha planta perenne da familia das asteráceas, distribuída ao longo de Europa. Ten interese farmacéutico debido as súas propiedades antiinflamatorias, derivadas de metabolitos secundarios, basicamente lactonas sesquiterpénicas. No estudo do presente caso, realizouse unha primeira aproximación xenómica como continuación de estudos xenéticos previos [1,2,3]. A partir de 120 individuos de 12 localidades do norte de España (Táboa 1), obtivéronse 5675 marcadores de polimorfismos de nucleótido único (SNPs, do inglés *Single Nucleotide Polymorphisms*). Dentro da batería de análises realizadas a partir destes marcadores atópanse análises de estrutura poboacional. Para estas, empregáronse dous paquetes de R, adegenet [4] e ParallelStructure [5], sitos nos repositorios, CRAN (https://cran.r-project.org/) e R-forge (https://r-forge.r-project.org/), respectivamente.

Localidade	Código	Nº Individuos
Ponte Pedrido (Guitiriz, Lugo)	PPED	10
Outeiro de Rei (Lugo)	OUT	12
Ponte de Bous (Serra dos Ancares, Lugo)	PBOU	12
Catro Carballos (Serra dos Ancares, León)	CCAR	5
Marco do Pozo (Serra dos Ancares, Lugo)	POZO	4
Alto do Couto (Serra do Courel, Lugo)	COUT	10
Pico Formigueiros (Serra do Courel, Lugo)	PFOR	9
Covadonga (Asturias)	COME	11
Boya (Zamora)	BOYA	13
Forcadura (Zamora)	FORC	14
Salduero (Vizcaia)	SALD	13
Ripollés (Girona)	RIPO	7

Táboa 1: Localidades co seu código asociado e o número de individuos mostrados.

O paquete adegenet permite inferir a estrutura xenómica dos nosos datos en dúas etapas principais Na primeira etapa definiríanse o número de unidades poboacionais máis probables segundo o criterio de información bayesiana (BIC, do inglés *Bayesian Information Criterion*) empregando a función "find.clusters". Posteriormente, na segunda etapa consistiría na realización de análises discriminantes de compoñentes principais (DAPCs, do inglés *Discriminant Analysis of Principal Components*) [6] coa función "dapc". Este método consta a súa vez de dous pasos: a análise de compoñentes principais ou PCA (*Principal Component Analysis*) + a análise discriminante ou DA (*Discriminant Analysis*). Os DAPCs teñen como obxectivo mostrar as diferencias entre grupos minimizando as variacións dentro dos grupos, de menor interese neste tipo de análises. A selección do número de compoñentes principais é unha decisión crítica, e existen diferentes criterios publicados na bibliografía científica.

Por outra banda, o paquete ParallelStructure permite a execución e a paralelización da ferramenta STRUCTURE [7]. Este programa, identifica o número de unidades poboacionais máis probable entre os analizados en base a certas asuncións xenéticas (e.g., equilibrio de Hardy-Weinberg dentro das unidades poboacionais identificadas) empregando unha aproximación bayesiana. Ao empregar de miles a millóns de marcadores, con longas iteracións (100,000-1,000,000), ademais de analizar un elevado número de unidades poboacionais posibles (Ks) para as nosas mostras, estas análises son computacionalmente moi intensivas. Isto implica que a paralelización (i.e., emprego paralelo de varias unidades de procesamento) é fundamental para poder manexar tempos de execución razoables tanto en local (e.g., ordenador de sobremesa) como en centros de supercomputación.

2. METODOLOXÍA

Co paquete de R adegenet utilizouse a función "find.clusters" testando até K = 20 unidades poboacionais ou clústeres, empregando 1,000,000 iteracións por cada K. Para a selección do número de clústeres máis probables seleccionáronse as Ks cos valores BIC máis baixos. Posteriormente, no DAPC, probáronse diferentes criterios para a selección do número de compoñentes principais :

(i) o criterio de porcentaxe de varianza conservada (capturando tantos PCs necesarios para reter tanto o 50% como para o 90%)

(ii) a-score, a diferenza entre a proporción de reatribucións con éxito de análise (discriminación observada) e os valores obtidos utilizando grupos aleatorios (discriminación aleatoria).

(iii) retención dun número de compoñentes principais equivalente ao número de Ks inferidos mediante "find.clusters" -1, empregando a configuración inicial das localidades, criterio baseado en [8].

Tras unha comparación de resultados, este último criterio foi o elixido para empregar en todas as análises de DAPC xa que ademais de reter unha elevada porcentaxe da variación total dos datos (i.e., >50%; primeiro criterio) os resultados foron similares.

As análises con ParallelStructure foron realizadas no supercomputador FinisTerrae III do Centro de Supercomputación de Galicia (CESGA). Testouse un número de K dende 1 ata 13 (número de localidades + 1) cos modelos ADMIXTURE (podería haber fluxo xénico e híbridos entre diferentes localidades) e frecuencias alélicas correlacionadas. Para a interpretación de resultados empregouse o programa web StructureSelector [9], que ofrece diferentes estimadores do valor de K máis probable [7, 10-11], e a representación visual empregando o programa CLUMPAK [12].

3. RESULTADOS

Os resultados atopados con ambos métodos foron consistentes entre si (Figura 1).

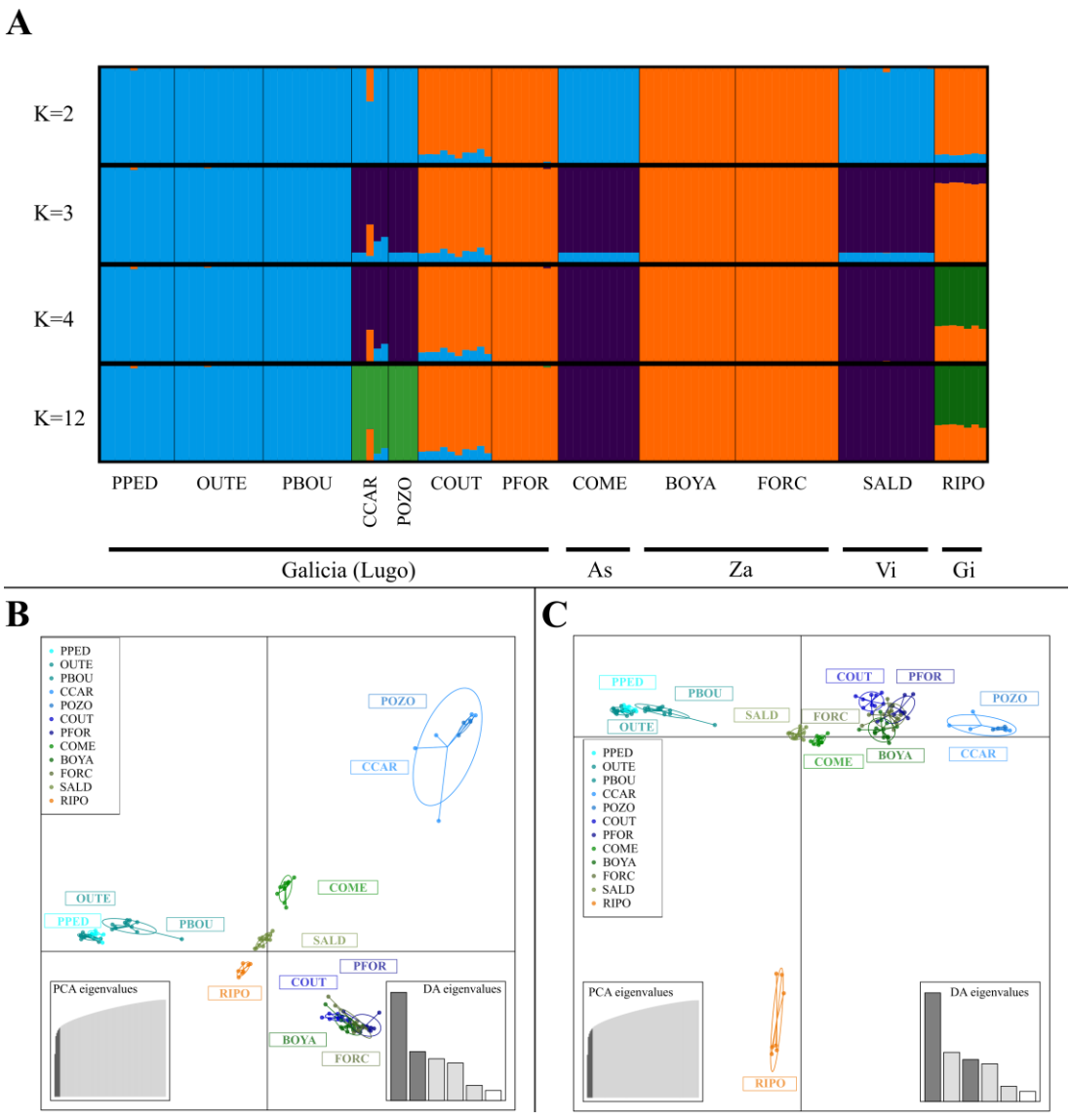


Figura 1: **A:** Visualización de CLUMPAK dos resultados de STRUCTURE para diferentes K (2,3 e 4 con apoio dos modelos). Cada individuo está representado como unha barra vertical dividida en segmentos segundo a proporción do xenoma pertencente a cada un dos clústeres inferidos (en cores distintas). Observamos dous grandes bloques, que ao aumentar K se diferencian en subgrupos. As: Asturias; Za: Zamora; Vi: Bizcaia; Gi: Xirona **B e C:** resultados de DAPC comparando o primeiro DA *eigenvalue* co segundo (B) e co terceiro (C) respectivamente. Os individuos están representados como puntos e as localidades como elipses de inercia.

Dependendo do estimador de K de STRUCTURE empregado obtivéronse distintos resultados. Isto apuntaría a unha xerarquía dos clústeres estruturais, en que algúns se subdividen en varios segundo consideramos a posibilidade de maiores K (Figura 1).

Os resultados suxerirían que hai dous clústeres principais, representados en azul e laranxa respectivamente (Figura 1) coincidindo con observacións de traballos previos. Ademais, obsérvase que localidades próximas entre si (CCAR-POZO) e afastadas (COME-SALD)agruparíanse en clústeres propios con maiores valores de K, suxerindo a implicación de factores demográficos e adaptación locais na estrutura poboacional.

AGRADECEMENTOS

Este traballo foi apoiado polos contratos da Deputación de Lugo para estudos de *Arnica montana* dos anos 2023 e 2024. Fernando Cabana foi beneficiario dunha Bolsa Iniciación á Investigación do Campus Terra (2024).Agradecemos o apoio bioinformático do CESGA.

Referencias

[1] Vera, M., Romero, R., Rodríguez Guitián, M., Barros, R., Real, C. e Bouza, C. (2015). Phylogeography and genetic variability of the *Arnica montana* chemotypes in NW Iberian Peninsula. *Silvae Genetica*, *63*, 293–300.

[2] Vera, M., Mora, G., Rodríguez-Guitián, M. A., Blanco, A., Casanova, A., Real, C., Romero, R. e Bouza, C. (2020). Living at the edge: population differentiation in endangered *Arnica montana* from NW Iberian Peninsula. *Plant Systematics and Evolution*, *306*, 44.

[3] Bouza, C., Lorenzo, I., Rodríguez-Guitián, M. A., Casanova, A., Real, C., Romero, R. e Vera, M. (2023). Genetic survey extension of the threatened Iberian *Arnica montana* L. revealed the presence of divergent plastid lineages and highly structured populations in northern Spain. *Genetic Resources and Crop Evolution*, *70*, 1677–1689.

[4] Jombart, T. e Ahmed, I. (2011). *adegenet 1.3-1*: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*, 3070–3071.

[5] Besnier, F. e Glover, K. A. (2013). ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers. *PLOS ONE*, 8, e70651.

[6] Jombart, T., Devillard, S. e Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94.

[7] Pritchard, J. K., Stephens, M. e Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*, 945–959.

[8] Thia, J. A. (2023). Guidelines for standardizing the application of discriminant analysis of principal components to genotype data. *Molecular Ecology Resources*, *23*, 523–538.

[9] Li, Y. L. e Liu, J. X. (2018). StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources*, *18*, 176–177.

[10] Evanno, G., Regnaut, S. e Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, *14*, 2611–2620.

[11] Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*, 608–627.

[12] Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. e Mayrose, I. (2015). CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*, 1179–1191.

TOPONOMASTICS, UMA FERRAMENTA PARA O ESTUDO DA TOPONÍMIA COM FOCO NO SUPERESTRATO

Afonso Xavier Canosa Rodrigues 1[1]

[1] IES Pedra da Auga (Ponteareas)

RESUMO

Obter listagens, mapas com a distribuição geográfica e processar os dados para a análise estatística é parte importante do estudo da toponímia. Apresentamos aqui uma série de funções de pesquisa e elaboração de informes (listagens, cartografia e gráficos) assim como exemplos de análises exploratórias obtidas com os scripts disponíveis no repositório Toponomastics, uma utilidade especialmente concebida para o trabalho com topónimos de superestrato de base antroponímica.

Palabras e frases chave: toponímia, antroponímia, objetos geográficos, entidades geográficas, R

1. INTRODUÇÃO

Toponomastics[1] é o nome dado a um repositório de funções desenhadas inicialmente para o estudo de temas (i.e. unidades lexicais em que segmentamos um topónimo) de origem antroponímica na toponímia. Os scripts permitem definir um tema (ou mais) e pesquisá-lo numa base de dados oferecendo listagens de resultados selecionados em função do tipo (ponto: localidade; ou polígono: freguesia ou concelho) assim como a sua distribuição espacial na área de estudo (mapas). Ainda que este tipo de perguntas pode ser respondido de modo direto ou com um mínimo de elaboração num SIG e mesmo em webs específicas de cartografia e toponímia, Toponomastics ambiciona criar um sistema em R que aproveite a versatilidade desta linguagem à hora de adicionar e processar os dados.

2. OBJETIVOS ESPECÍFICOS

Os scripts atuais de Topomastics foram concebidos para operarem com as bases de dados do Plan Visor Básico da Xunta[2], por estarmos a trabalhar nesta área e entendermos que oferece dados fiáveis a nível toponímico, para além de prover capas com os diferentes níveis administrativos. Porém, os dataframes são modificáveis e ampliáveis, e de facto estamos já a incorporar dados do Norte de Portugal, um continuum necessário na matéria de estudo. Objetivo geral de Toponomastics é ser aplicável a bases de dados doutras partes de Europa relevantes para o objeto principal de pesquisa: topónimos de base antroponímica e morfologia ditemática (dous temas num único topónimo).

Topomastics está orientado para a análise de dados toponímicos com objetivos linguísticos. Nesta primeira fase o objetivo concreto é obter e processar os dados para oferecer respostas e produzir informes exploratórios rápidos (listagens, mapas, gráficos). A sequência dum script tipo segue um princípio cíclico[1] para resolver perguntas como: que topónimos contêm o tema x? qual é a distribuição espacial do tema x? qual é distribuição espacial do tema x vs. tema y?

[1] https://github.com/afonsoxavier/toponomastics
[2] https://mapas.xunta.gal/visores/pba

3. ESTRUTURA DUM SCRIPT TIPO

3.1 Nos scripts de exemplo oferecidos no repositório, os dados podem ser carregados automaticamente. Por segurança e para evitar a repetição de descarregamentos desnecessários, a função tem de ser ativada no próprio script. Uma vez obtidos os dados, selecionamos os mais relevantes dentro do conjunto, homogeneizamos colunas equivalentes mas com diferentes nomes nas taxonomias de origem para evitar a acumulação de NAs e melhorar a operatividade e finalmente ordenamos as variáveis mais relevantes. Toponomastics trabalha inicialmente com ficheiros tipo *shape* que converte num dataframe com o pacote sf[2], o utilizado de modo preferente para operar com os objetos geográficos.

3.2 Uma vez preparados os dataframes, a aplicação mais recorrente é a de pesquisa. A função *data_search* pesquisa um tema que nesta altura requer REGEX para precisar a posição prototemática ou deuterotemática (ambos casos são contemplados nos scripts de exemplo) e permite especificar o tipo de entidade (localidade, freguesia, concelho, comarca ou todas à vez). Da sua parte, *search_comarca* agiliza o processo de pesquisa sobre uma ou várias comarcas. A função *unique_toponym* produz um único resultado quando o topónimo se repete em vários níveis de tipo de entidade num mesmo espaço (ex. um concelho que também é freguesia ou lugar) para evitar distorções nas análises frequentísticas. Finalmente, *entropy* devolve a entropia do sistema gerado pelos topónimos e entidades geográficas associadas em que ocorre um tema.

3.3. Toponomastics oferece diferentes tipos de resultados sobre as pesquisas. A função *list_toponimos,* cria uma listagem ordenada com todos os topónimos para um tema (e posição). Cada entidade geográfica associada ao topónimo vem acompanhada da entidade maior a que pertence. Uma segunda função, *barplot_freq*, apresenta um gráfico de barras com as frequências dos topónimos a partir do tema dado e *barplot_freq_entropy* adiciona o número de expressões toponímicas, entidades geográficas e entropia como subtítulo. Finalmente *full_report* pesquisa, cria uma listagem de entidades e representa os resultados num mapa, todo na mesma função.

4. EXEMPLOS DE ANÁLISES EXPLORATÓRIAS

Os casos a seguir foram obtidos com os scripts de mostra oferecidos no repositório de Toponomastics[3]. No primeiro exemplo utilizamos uma função de pesquisa para o deuterotema -ufe, produzimos listagens, filtramos entidades concorrentes para um mesmo topónimo, representamos no mapa as entidades geográficas e finalmente obtemos um gráfico com a produtividade de cada um dos topónimos associados ao tema (fig.1).

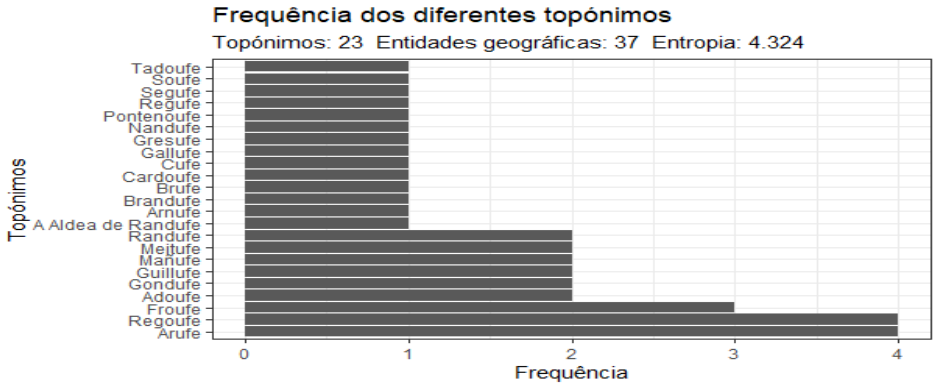


Figura 1 Listagem e frequências dos topónimos com deuterotema -ufe

[3] Visor_toponomastics_main.R exemplo dum script tipo, experiment_maps.R para exemplos de mapas e visor_toponomastics_particular_area.R, estudo duma área geográfica mais definida.

Para além de vermos a sua presença geográfica, podemos comprovar se há alguma relação espacial com outras formas. Neste caso (fig. 2) observamos a tendência à complementaridade que evidencia as formas pesquisadas serem variantes dum mesmo tema com distribuição geográfica.

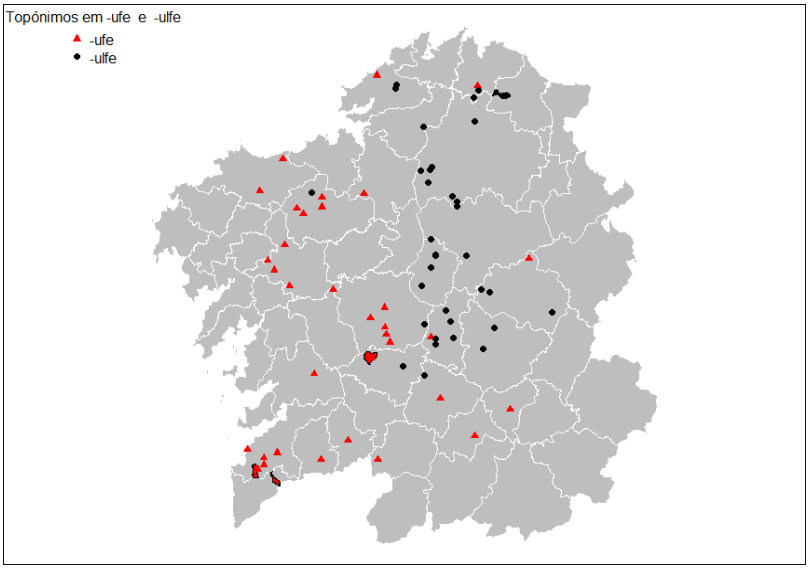


Figura 2 Relevância espacial da distribuição do deuterotema -ufe e variante -ulfe

Se observamos que há uma relação local relevante podemos focar numa área. No seguinte exemplo vemos a distribuição das formas -mil e -mir (fig. 3).

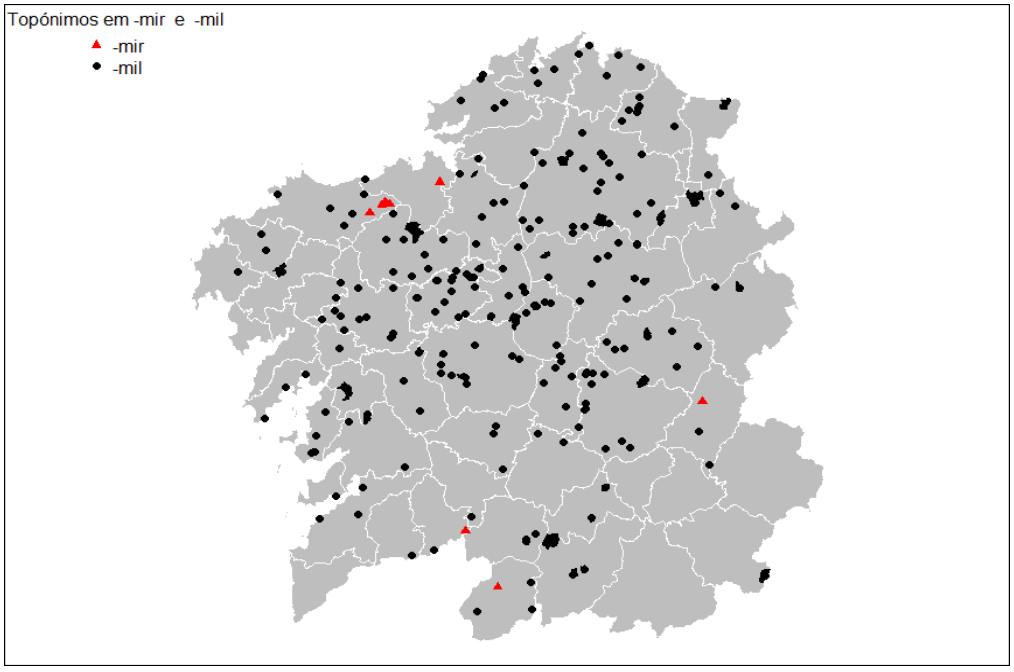


Figura 3 Distribuição das formas -mir e -mil

A presença dum clúster para as formas -mir faz com que exploremos com mais pormenor a área, por exemplo com um mapa que restrinja a pesquisa a nível comarcal (fig. 4).

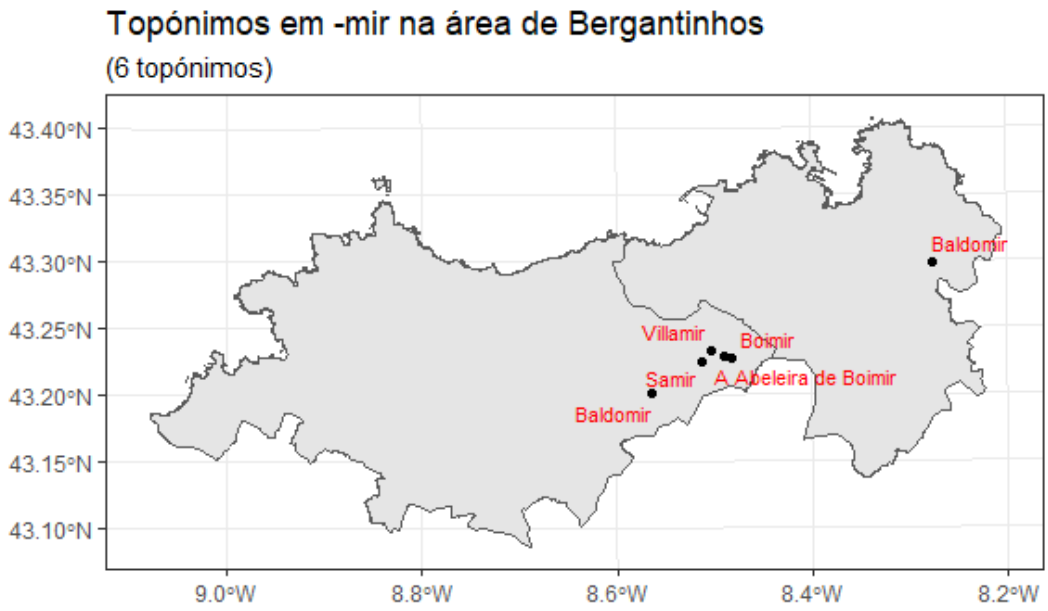


Figura 4 Clúster das formas -mir e -mil na área de influência de Bergantinhos

5. RESULTADOS E CONCLUSÔES

Ainda que em fase inicial e com necessidade de desenvolvimento, Toponomastics reúne já uma série de funções capazes de oferecerem resultados relevantes no trabalho com dados toponímicos: listagens, distribuição espacial e análise estatística exploratória. No seu estado atual, porém, requer conhecimentos avançados da linguagem R para criar novos scripts e adaptar as funções a novas bases de dados. O trabalho futuro contempla ampliar as análises de tipo geográfico e linguístico, assim como desenvolver uma interface de modo que a ferramenta opere sem necessidade de manipular o código. Com esta comunicação pretendemos sobretudo chamar a atenção sobre a potencialidade de R para a análise de dados toponímicos. Há, sem dúvida, soluções SIG com uma produção gráfica incomparável e uma grande capacidade para a análise espacial. Contudo, R continua a ser um instrumento que nos permite obter e tratar os dados com facilidade e, quando de dados espaciais se tratar, oferece uns resultados mais do que aceitáveis para as análises exploratórias que, no nosso caso, são as pretendidas. Onde pouco conhecemos ainda, começamos explorando.

Referências

[1] Peng R., Matsui E. (2015). *The Art of Data Science*. Leanpub [sem lugar de publicação]

[2] Pebesma E., Bivand R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC, https://r-spatial.org/book

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

NON METAS A GAMBA! ANÁLISES DE DATOS XENÓMICOS CON R

Adrián Casanova[1], Miguel Hermida[1], Paulino Martínez[1], Inmaculada Carrasco[2], Francisca Robles[3], Rafael Navajas[3], Roberto de la Herrán[3], Carmelo Ruiz[3]

[1] Departamento de Zooloxía, Xenética e Antropoloxía Física, Facultade de Veterinaria, Campus Terra, Universidade de Santiago de Compostela, Lugo, España
[2] Asoc. Organización de Productores Pesqueros de Motril OPP85, Motril (Granada), España
[3] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, España

RESUMO

O entorno de R e o seu amplo catálogo de paquetes permiten traballar cun gran volume de datos xenómicos dun xeito reproducible, sistemático e computacionalmente eficiente. Neste caso de estudo móstrase a aplicación de R, xunto con outros programas libres e gratuítos, no procesamento bioinformático dos primeiros datos xenómicos dun crustáceo mariño cun elevado interese comercial en España: *Plesionika edwardsii* (Brandt, 1851), coñecido comercialmente como camarón soldado ou quisquilla. Os resultados obtidos poderán contribuír á xestión e valorización comercial dos caladoiros pesqueiros así como á conservación das poboacións naturais.

Palabras e frases chave: *Plesionika edwardsii*, quisquilla, xenómica, pesca, Mediterráneo occidental.

1. INTRODUCIÓN

O camarón soldado ou quisquilla (*P. edwardsii*) é un crustáceo decápodo mariño cunha coloración alaranxada, coa presencia de varias raias lonxitudinais máis escuras na parte dorsal do abdome (Figura 1). As femias adultas caracterízanse por presentar ovos azuis. Este crustáceo pode acadar até uns 17 centímetros de lonxitude [1]. A quisquilla habita frecuentemente nun rango de profundidade entre 200-500 metros, tratándose dunha especie nectobentónica, ligada ao fondo mariño pero con migracións verticais a traveso da columna de auga [2]. Presenta unha distribución circunglobal, sendo un importante recurso pesqueiro no Mediterráneo occidental. No ano 2022, o total de capturas en España ascenderon a case 400 toneladas de peso vivo [3], cun valor medio nas lonxas duns 17 euros/kg [4,5].

Este traballo centrouse principalmente en mostras do mar de Alborán (Mediterráneo occidental), situadas en diferentes secos. Estes son montes submarinos que se agrupan en extensas mesetas, hábitats para unha multitude de especies de elevado valor ecolóxico e económico (caladoiros de pesca). Debido á localización en diferentes secos, aos diferentes réximes hidrográficos e as correntes mariñas existentes as poboacións do mar de Alborán poderían presentar un certo grao de illamento xenético. Isto podería aportar características específicas ás poboacións desta rexión que puideran ter interese comercial.

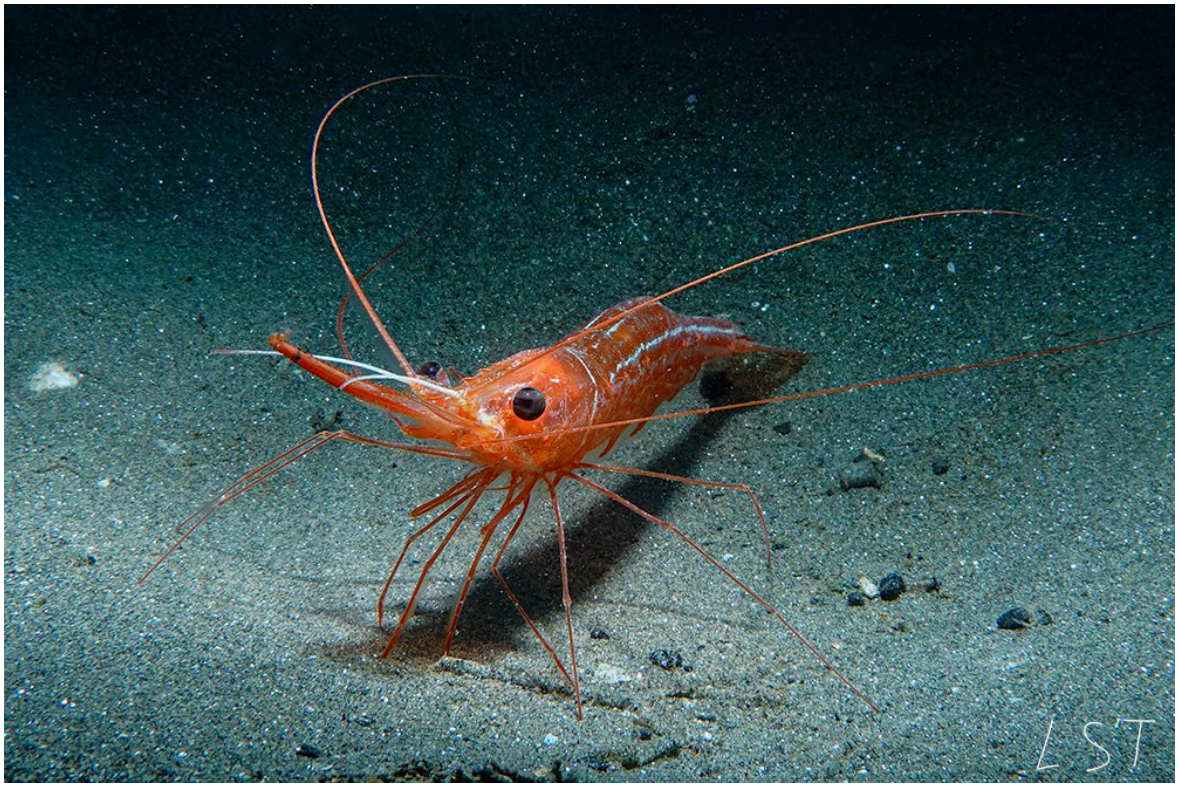


Figura 1: Exemplar de quisquilla (*P.edwarsii*). Imaxe de Luis Sánchez Tocino (Universidad de Granada).

2. METODOLOXÍA

Un total de 128 mostras procedentes de dúas localidades situadas no Océano Atlántico (Illas Canarias e Huelva; Táboa 1) e do mar de Alborán (Mediterráneo Occidental) foron empregadas para obter un panel de miles de marcadores moleculares de tipo SNP (i.e., polimorfismos no ADN de nucleótido único) localizados nun xenoma previamente secuenciado de *P.edwardsii*.

Localidade	Código	Océano/mar	N mostras
Illas Canarias	CA	Atlántico	28
Huelva	HU	Atlántico	24
Seco de Motril	SM	Mediterráneo	24
Seco de los Olivos	SO	Mediterráneo	28
Seco de Torrox	ST	Mediterráneo	24

Táboa 1: Mostras de quisquilla incluídas no presente traballo.

O panel bruto de SNPs foi filtrado por calidade, empregando diferente software bioinformático, incluídos os paquetes de R {**radiator**} [6] e {**genepopedit**} [7] que asemade facilitan o ordenamento das mostras e a interconversión de formatos en arquivos moi pesados. A partir do panel filtrado de SNPs realizáronse unha serie de análises con diferentes paquetes de R, tanto xenéricos {**base**} [8] como específicos deste tipo de aproximacións xenómicas {**dartR**} [9,10]. Para estimar a diversidade xenética, a materia prima da evolución das diferentes poboacións naturais, empregáronse os paquetes de R {**diveRsity**} [11] e {**genepop**} [12]. Para avaliar a diferenciación xenética entre localidades (F_{ST}; cun rango de 0 a 1) empregouse {**genepop**} e {**StAMPP**} [13] (Caixa de texto 1). Para o estudo da estrutura poboacional empregouse {**adegenet**} [14-16] e o software STRUCTURE [17], a través do paquete {**ParallelStructure**} [18] que permite paralelizar o traballo en múltiples cores. Varias análises con R realizáronse no Centro de Supercomputación de Galicia (CESGA) trala creación previa dunha libraría persoal.

```
# Ler o arquivo cos xenotipados en formato genepop usando o paquete {adegenet}
genind<-read.genepop("N128b_6x_360_DEF.recode.gen",ncode = 3L,quiet = FALSE)

# Conversión dun obxecto genind nun obxecto genlight {dartR}
genlight<-gi2gl(gi = genind,parallel = TRUE,verbose = NULL)

# Estimación da FST por pares de localidades {StAMPP}
output<-stamppFst(genlight,nboots = 10000,percent = 95,nclusters = 20)

# Exportación dos resultados {base}
write.matrix(output$Fsts,file = "matrix_Fst_B10000_n128_Plesionika.txt")
write.matrix(output$Pvalues,file = "matrix_Pvalues_B10000_n128_Plesionika.txt")

out_Boots<-output$Bootstraps
out_Boots_t<-t(out_Boots)
write.table(out_Boots_t,file = "boots_10000_n128_Plesionika.txt")
```

Caixa de texto 1: Liñas do *script* de R para estimar a estrutura poboacional (F_{ST}) co paquete {**StAMPP**}.

3. R-ESULTADOS

Tralos pasos de filtrado por calidade, retivéronse un total de 17.416 SNPs localizados ao longo do xenoma de *P.edwardsii*. A diversidade xenética foi moi similar entre as diferentes localidades (He = 0.168-0.174). Os datos obtidos non mostraron diferenciación xenética entras as localidades do Mar de Alborán, suxerindo un elevado fluxo xénico entre elas, pero si entre o Mar de Alborán e as dúas localidades atlánticas (Figura 2). A maior diferenciación xenética produciuse entre as dúas localidades atlánticas (Huelva e Illas Canarias; F_{ST} = 0,016; Táboa 2).

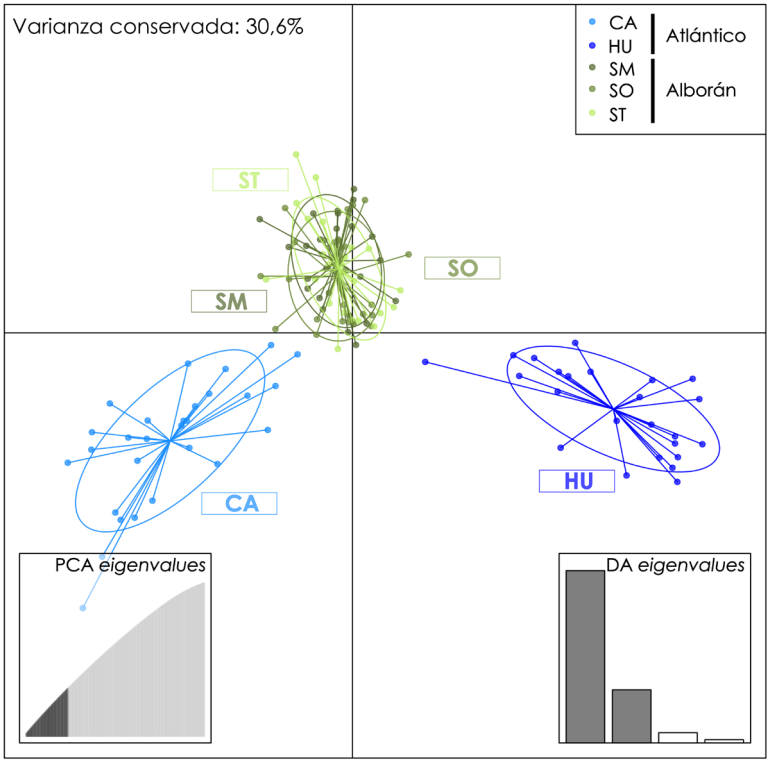


Figura 2: Gráfico de dispersión das quisquillas (N = 128) nos dous *eigenvalues* principais da Análise Discriminante do DAPC [15] (*Discriminant Analysis of Principal Components*). As localidades están coloreadas en tonalidades azuis (Océano Atlántico) e verde (Mar de Alborán). Este gráfico representa os individuos como puntos e as localidades coma elipses de inercia. DAPC realizada co paquete de R {**adegenet**} e as súas dependencias.

	CA	HU	SM	SO	ST
CA	—	***	***	***	***
HU	0,016	—	***	***	***
SM	0,008	0,006	—	*	NS
SO	0,008	0,006	0,001	—	NS
ST	0,008	0,006	0,000	0,000	—

Táboa 2: Matriz cos valores de diferenciación xenética (F_{ST}) por pares de localidades (debaixo da diagonal) e os limiares de significación estatística (enriba da diagonal). *** p-valor < 0,001, ** p-valor < 0,01, * p-valor < 0,05, NS Non Significativo. Estimacións obtidas co paquete de R {**StAMPP**}.

4. CONCLUSIÓNS

As localidades de quisquilla do Mar de Alborán poderían ser incluídas na mesma Unidade de Xestión (MU, *Management Unit*). Con todo, a diferenciación xenética coas localidades atlánticas foi moi baixa, especialmente coa das Illas Canarias a pesares de estar a máis de 1.000 quilómetros de distancia. É posible que nas diferentes características organolépticas que a quisquilla poida presentar haxa un maior peso das variables ambientais respecto ás variables xenómicas.

AGRADECEMENTOS

Este traballo realizouse no marco do proxecto "Análisis genético de *Plesionika edwardsii* en poblaciones del Mar de Alborán (PLESIGEN)" B-BIO-678-UGR20, financiado polos Proyectos I+D+i del Programa Operativo FEDER Andalucía 2020 de la Consejería de Transformación Económica. Industria, Conocimiento y Universidades de la Junta de Andalucía. Adrián Casanova é financiado por un contrato posdoutoral Xunta de Galicia-Campus Terra (2022). Agradecemos o apoio bioinformático do CESGA.

Referencias

[1] ICTIOTERM. (2024). *Plesionika edwardsii* (Brandt, 1851). *ICTIO-TERM. Base de datos terminológicos y de identificación de especies pesqueras de las costas andaluzas.* Dispoñible en: http://www.ictioterm.es/nombre_cientifico.php?nc=231 (data de acceso: 3 de setembro de 2024).

[2] Company, B.J, Sardà, F. (1997). Reproductive patterns and population characteristics in five deep-water pandalid shrimps in the Western Mediterranean along a depth gradient (150-1100 m). *Marine Ecology Progress Series* 148, 49-58.

[3] FAO. (2023). Global capture production Quantity (1950 - 2022). *Food and Agriculture Organization of the United Nations Fisheries and aquaculture Statistical Query Panel*. Dispoñible en: https://www.fao.org/fishery/statistics-query/ (data de acceso: 3 de setembro de 2024).

[4] Lillo-Bañuls, A., Fuster, B., Merino, F., Mora, J., Ortiz-Pérez, S. (2024). Estudio socioeconómico del sector pesquero de la Comunidad Valenciana. *Universidad de Alicante*.

[5] IDAPES. (2024). Consultas estadísticas pesqueras. Primera venta de pesca fresca en lonja. *Consejería de Agricultura, Pesca, Agua y Desarrollo Rural*. Dispoñible en: https://www.juntadeandalucia.es/agriculturaypesca/idapes/ (data de acceso: 3 de setembro de 2024).

As referencias do software pódense consultar usando este QR

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Análise de datos contables mediante R. Exemplo de análise de certos datos de facturación e gastos na Xunta de Galicia

Marcos Fernández Arias[1]

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

RESUMO

Caso práctico de exemplo de análise, realizado recentemente, de certos datos contables (precios, contratos, facturación...) de provedores da Xunta de Galicia.

Palabras e frases chave: tidyverse, business reports, data enrichment

1. INTRODUCIÓN

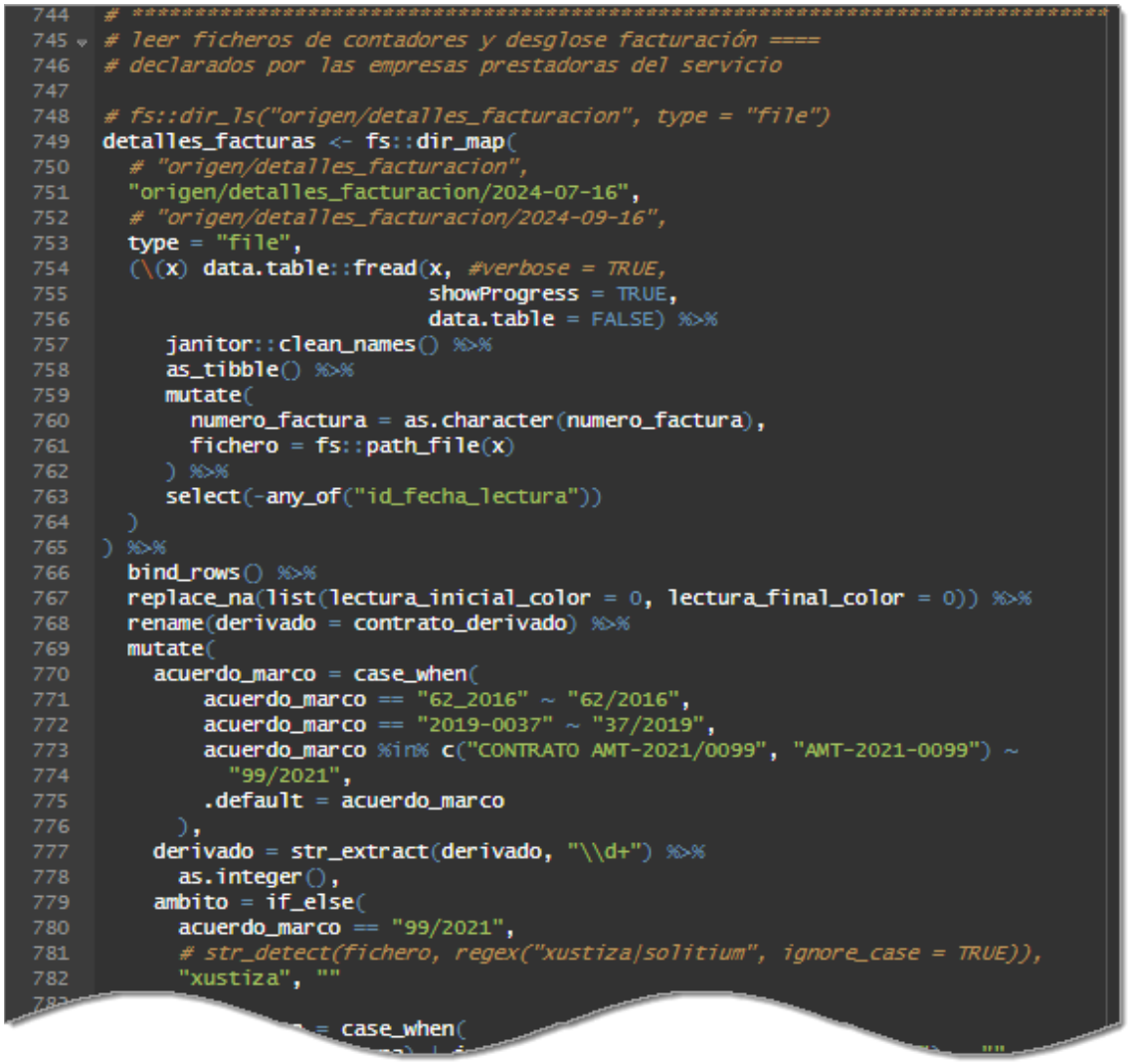
Explicación práctica dos métodos aplicados recentemente para resolver varios problemas que xurdiron ao intentar cruzar e analizar certos conxuntos de datos relativos á contabilidade e facturación (tamaño: millóns de filas) relativos a contratos, facturación e medicións de contadores de uso.

2. Procesamento dun ficheiro CSV que contén erros de formato e non resulta lexible

```r
# pacman::p_load(archive)
snmp_0 <- archive::archive_read(
  "origen/SNMP/snmp-2024-09.7z",
  file = "snmp-2024-09.txt") %>%
  read_csv(col_names = FALSE) %>%
  set_names(
    c("fecha", "hora", "ip", "name", "serial", "model", "c_bn", "c_color")
  ) %>%
  rowid_to_column() %>%
  filter(!is.na(fecha) & !is.na(serial)) %>%
  mutate(
    fecha = ymd(fecha),
    hora = as.integer(hora),
    # c_bn = as.integer(c_bn),
    # c_color = as.integer(c_color)
  )

lines <- archive::archive_read(
    "origen/SNMP/snmp-2024-09.7z",
    file = "snmp-2024-09.txt") %>%
  read_lines() %>%
  as_tibble_col(column_name = "txt") %>%
  rowid_to_column()
lines

# líneas que no contienen datos válidos sino un mensaje de error ====
lines %>%
  slice(setdiff(1:nrow(lines), snmp_0$rowid)) %>%
  count(txt)


# líneas que no se han parseado bien por contener comas dentro de strings ====
rowids_necesario_repetir_parseo <-
  snmp_0 %>%
  filter(!is.na(c_color) & is.na(parse_integer(c_color))) %>%
  pull(rowid)
length(rowids_necesario_repetir_parseo)
rowids_necesario_repetir_parseo
```

Explicaremos un caso de procesar un ficheiro CSV de medicións (de máis de un millón de filas) que contiña erros de formato: non era un ficheiro CSV válido senón un TXT que case era un CSV.
Non podía ser lido directamente mediante ningunha librería, polo que foi necesario programar un procesamento a medida un pouco máis avanzado.

3. Procesamento de múltiples ficheiros CSV co mesmo formato



```
# ********************************************************************************
# leer ficheros de contadores y desglose facturación ====
# declarados por las empresas prestadoras del servicio

# fs::dir_ls("origen/detalles_facturacion", type = "file")
detalles_facturas <- fs::dir_map(
  # "origen/detalles_facturacion",
  "origen/detalles_facturacion/2024-07-16",
  # "origen/detalles_facturacion/2024-09-16",
  type = "file",
  (\(x) data.table::fread(x, #verbose = TRUE,
                          showProgress = TRUE,
                          data.table = FALSE) %>%
     janitor::clean_names() %>%
     as_tibble() %>%
     mutate(
       numero_factura = as.character(numero_factura),
       fichero = fs::path_file(x)
     ) %>%
     select(-any_of("id_fecha_lectura"))
  )
) %>%
  bind_rows() %>%
  replace_na(list(lectura_inicial_color = 0, lectura_final_color = 0)) %>%
  rename(derivado = contrato_derivado) %>%
  mutate(
    acuerdo_marco = case_when(
        acuerdo_marco == "62_2016" ~ "62/2016",
        acuerdo_marco == "2019-0037" ~ "37/2019",
        acuerdo_marco %in% c("CONTRATO AMT-2021/0099", "AMT-2021-0099") ~
          "99/2021",
        .default = acuerdo_marco
      ),
    derivado = str_extract(derivado, "\\d+") %>%
      as.integer(),
    ambito = if_else(
      acuerdo_marco == "99/2021",
      # str_detect(fichero, regex("xustiza|solitium", ignore_case = TRUE)),
      "xustiza", ""
                     = case_when(
```

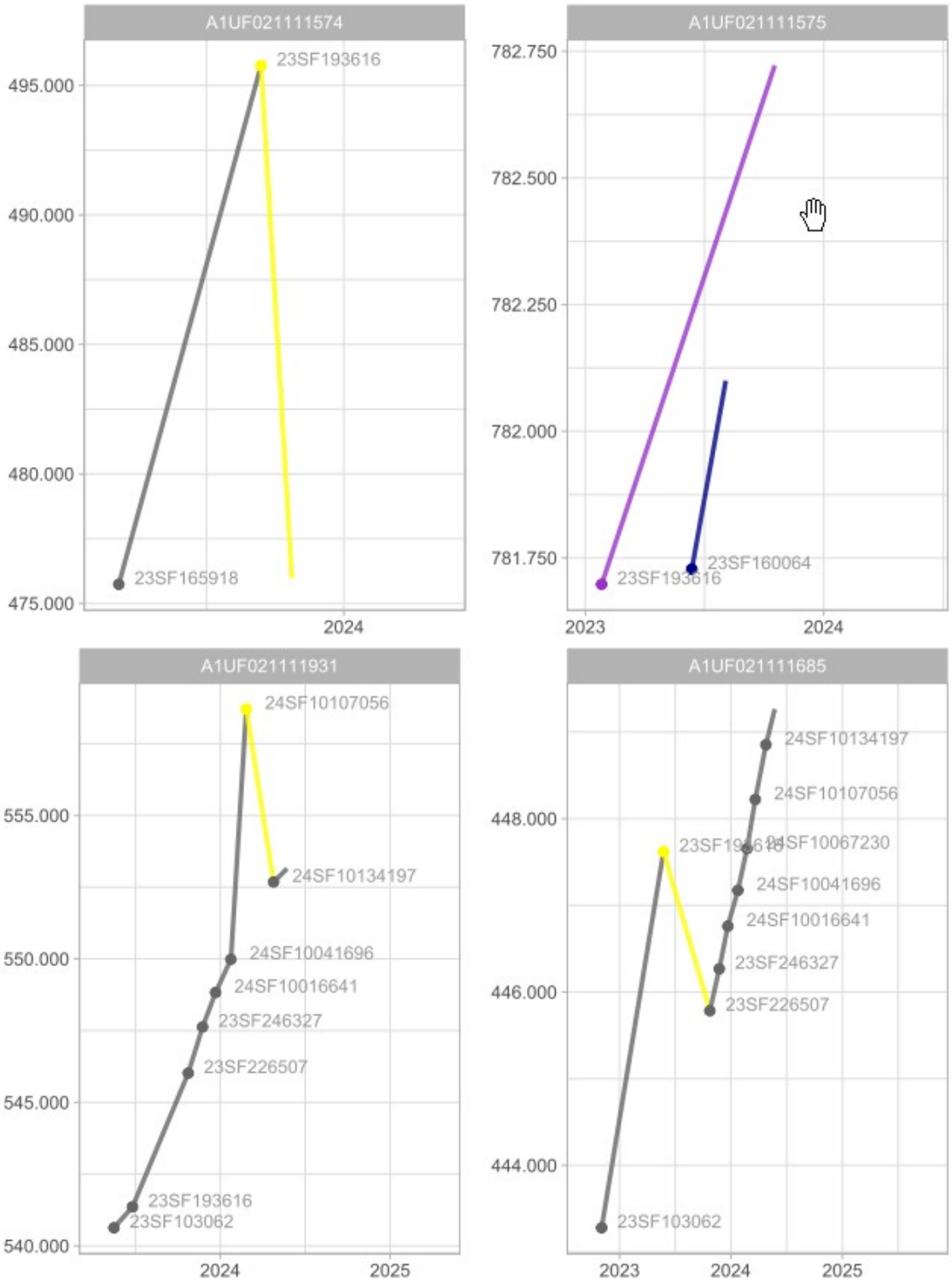
Mediante programación funcional (paquete *purrr* e paquete *fs*) conseguimos leer decenas de ficheiros CSV de similar formato.
Cada ficheiro contén os datos dun período de tempo concreto. O proceso consolida varios ficheiros nun único dataframe (tibble).

3. Xeración de gráficos sofisticados de apoio ás explicacións numéricas

```
g_1 <- datos %>%
  filter(!is.na(bn)) %>%
  mutate(
    lectura_final_bn = if_else(lectura_inicial_bn == lectura_final_bn,
                               lectura_final_bn + 1, lectura_final_bn),
    bn_category = case_when(
      eval_bn == "" ~ "gray40",
      eval_bn == "-" ~ "yellow",
      eval_bn == ".<" ~ "darkgreen",
      eval_bn == ".>" ~ "darkblue",
      eval_bn == "<." ~ "green",
      eval_bn == ">." ~ "blue",
      eval_bn == "rectif" ~ "darkorchid",
      eval_bn == "dup" ~ "red",
      .default = "pink"
    )
  ) %>%
  ggplot(aes(y = fecha_lectura_inicio, x = lectura_inicial_bn)) +
  geom_segment(aes(xend = lectura_final_bn,
                   yend = fecha_lectura_fin,
                   color = bn_category),
               alpha = .8,
               linewidth = 1.1) +
  geom_point(aes(color = bn_category), size = 2) +
  geom_text(aes(label = numero_factura),
            hjust = -.15, vjust = 0,
            # angle = 30,
            size = 3) +
  scale_x_continuous(
    labels = label_comma(big.mark = ".", decimal.mark = ","),
    # minor_breaks = scales::minor_breaks_width(1000, 0),
    # minor_breaks = scales::minor_breaks_n(1),
  ) +
  scale_y_date(
    date_labels = "%Y",
    breaks = "1 years",
    minor_breaks = "1 months",
    # minor_breaks = scales::minor_breaks_width("6 months", 0),
    expand = expansion(mult = c(.2, 1))
  ) +
  labs(
    # title = "irregularidades en contador de páginas en bn",
    x = "contador bn",
    y = NULL
  ) +
  theme(
    panel.grid.minor = element_line(
      linetype = 'dashed', linewidth = 0.1, color = 'gray90')
  ) +
  scale_colour_identity(guide = "none") +
  coord_flip() +
  aspect.ratio = 1.5)
```

Mediante *ggplot* automatizamos a creación de informes PDF de varias páxinas, con deceas de gráficos por páxina.
Utilizamos *ggforce::facet_wrap_paginate* e *ggplot2::ggsave*.

irregularidades en contador de páginas en bn



3. Xeración de recomendacións mediante fórmulas

```r
# redistribución de uso de impresoras en color dentro de una localización =====
redistribucion_uso_color <- prorrateado_segun_impresora %>%
  rename_with(~ str_remove(., "^media_")) %>%
  left_join(
    contratos %>%
      select(numero_serie, localizacion, salida, precio_bn, precio_color),
    by = "numero_serie"
  ) %>%
  left_join(
    cmdb %>%
      select(numero_serie, localizationtag, support_group,
             conselleria, model, note, owner),
    by = "numero_serie"
  ) %>%
  # count(localizacion, sort = TRUE)
  # count(localizationtag, sort = TRUE)
  filter(!is.na(localizationtag)) %>%
  filter((is.na(salida) | salida == "color") & color_mes > 0) %>%
  # filter(!is.na(note)) %>% pull(note)
  arrange(precio_color, desc(color_mes)) %>%
  group_by(conselleria, localizationtag, support_group) %>%
  filter(n() > 1) %>%
  summarise(
    n_impresoras = n(),
    # color_mes2 = str_flatten_comma(color_mes),
    # gt = first(precio_color) < nth(precio_color, 2),
    total_color_mes = str_c(
      str_c(color_mes, collapse = " + "), " = ", sum(color_mes, na.rm = TRUE)
      ),
    precios_color = str_flatten_comma(precio_color),
    coste_color_mes = sum(coste_color_mes) * 1.21,
    coste_ideal = min(precio_color) * sum(color_mes) * 1.21,
    ahorro_potencial = coste_color_mes - coste_ideal,
    impresoras_minimo_precio =
        if_else(precio_color == min(precio_color),
                str_c(model, " (", numero_serie, ")"),
                "") %>%
      str_c(collapse = "_") %>%
      str_remove("_*$") %>%
      str_remove("^_*") %>%
      str_replace_all("_", ", ") %>%
      str_squish(),
    texto_recomendacion = str_c(
      "Para impresión en color, recomendamos utilizar preferentemente ",
      if_else(
        sum(precio_color == min(precio_color)) > 1,
        "las impresoras ", "la impresora "
      ),
      # first(model),
      impresoras_minimo_precio,
      " en lugar de las otras por tener un precio por página menor (",
      first(precio_color), " €/página).",
      " Con lo que se podría llegar a conseguir un ahorro de hasta ",
      round(ahorro_potencial), " €/mes."
    ),
    .groups = "keep"
    # .by = c(support_group, localizationtag)
  ) %>%
  filter(ahorro_potencial > 2) %>%
  arrange(desc(ahorro_potencial))
redistribucion_uso_color$texto_recomendacion %>%
```

Tamén xeramos de maneira automatizada (mediante R) decenas de recomendacións de uso a medida (para reducir gastos, dirixidas aos usuarios finais).

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Aplicativos para Aprendizagem Acelerada das Iterações do Método Simplex

Luciane Ferreira Alcoforado - Academia da Força Aérea Brasileira (AFA)

RESUMO

Este trabalho apresenta a criação de dois aplicativos Shiny desenvolvidos para facilitar o ensino do método Simplex em programação linear. Esses aplicativos permitem a manipulação e resolução de problemas de programação linear de forma mais eficiente, automatizando a construção e operação de matrizes e vetores no procedimento iterativo do método. O objetivo é melhorar a compreensão dos conceitos fundamentais do método Simplex, acelerando o tempo gasto nos cálculos matriciais e promovendo uma experiência de aprendizado mais dinâmica.

Palavras chave: Método Simplex, iterações, operações matriciais, aplicativo shiny.

1. INTRODUÇÃO

O ensino do método Simplex em programação linear frequentemente enfrenta desafios relacionados à manipulação de elementos matriciais. Estudantes muitas vezes encontram dificuldades na construção e operação de matrizes e vetores, o que pode limitar a exploração de diferentes problemas devido ao tempo necessário para estruturar e calcular as iterações do algoritmo Simplex.

> "Diretamente relacionadas com o cálculo matricial, destacam-se as dificuldades relacionadas com a multiplicação de matrizes, facto explicável porque o algoritmo da multiplicação de matrizes é novo para os alunos e diferencia-se da habitual multiplicação de números reais." (Barros, Araújo e Fernandes, 2013) [5] .

Além disso, o trabalho com os conteúdos de Matrizes, Determinantes e Sistemas Lineares (MDSL) apresenta dificuldades adicionais, pois são estruturas extensas e que necessitam de muitas operações aritméticas precisas para serem realizadas.

> "A multiplicação de matrizes, embora possível manualmente para elementos com baixa dimensão, torna-se virtualmente impraticável de ser realizada, em sala de aula, quando a ordem da matriz é superior a três. O excesso de cálculos não contribui para manter o interesse dos alunos, principalmente daqueles que apresentam dificuldades e que se constituem na maioria dos alunos." (Steinhorst, 2011) [6] .

Essas dificuldades são corroboradas por estudos que mostram que a complexidade das operações matriciais pode ser um obstáculo significativo no aprendizado de programação linear. A necessidade de realizar cálculos precisos e a estruturação de matrizes de alta ordem consomem tempo e podem desmotivar os alunos, limitando a prática e a exploração de diferentes cenários de problemas.
O algoritmo Simplex, desenvolvido por George Dantzig, é uma técnica algorítmica utilizada para encontrar a solução ótima de problemas de programação linear [4]. Ele envolve a manipulação de matrizes e vetores para iterativamente melhorar a solução até que a solução ótima seja encontrada. O método Simplex utiliza um estrutura matricial que representa o sistema de equações lineares do problema. Cada iteração do algoritmo envolve operações matriciais, como transposição, inversão e multiplicação de matrizes.
Para enfrentar esses desafios, este artigo apresenta dois aplicativos Shiny desenvolvidos para acelerar o processo de manipulação e resolução de problemas de programação linear. Esses aplicativos permitem que os alunos tenham acesso a uma ampla variedade de problemas, com diferentes números de variáveis e restrições, e possam realizar as iterações do algoritmo até atingir sua regra de parada e analisar a solução final. O objetivo é ilustrar como a linguagem R pode ser utilizada para inovar no aprendizado do método simplex através de um aplicativo funcional.

Ao automatizar a construção e manipulação de matrizes, os aplicativos criados não apenas economizam tempo, mas também permitem que os alunos se concentrem na compreensão dos conceitos fundamentais do método Simplex, em vez de se perderem em cálculos tediosos. Dessa forma, a experiência de aprendizado se torna mais rica e envolvente, promovendo uma melhor assimilação dos conteúdos e uma maior motivação para explorar diferentes problemas de programação linear.

2. APRESENTAÇÃO DOS APLICATIVOS

A base teórica para a construção dos aplicativos considera o problema que expressaremos na forma matricial, sendo A_0 uma matriz $m \times n'$ que representa os coeficientes das restrições do modelo inicial; c_0 que representa o vetor custo, x_0 que representa o vetor das variáveis de decisão do modelo inicial e b que representa o vetor de recursos, cujas dimensões são respectivamente $n' \times 1$, $n' \times 1$ e $m \times 1$. Assim, o modelo matemático inicia com a seguinte estrutura:

$$\text{Função Objetivo: } min \text{ ou } max \ z = c_0^T \cdot x_0$$
$$\textit{Sujeito a: } A_0 \cdot x_0 \leq b,\ x_0 \geq 0$$

Após, coloca-se o problema na **forma padrão** em que a função objetivo passa a ser de minimização (se inicialmente era max, multiplica-se o vetor c por -1) e são acrescentadas a cada restrição do tipo $\leq$, m novas variáveis de folga $x_{n'+1}, ...x_{n'+m}$, totalizando $n = n' + m$ variáveis, alterando assim as dimensões dos elementos matriciais e considerando o problema na forma de minimização com restrições de igualdade. O modelo forma padrão se apresenta com a seguinte estrutura:

$$min \ z = c^T \cdot x$$
$$\textit{Sujeito a: } A \cdot x = b,\ x \geq 0 \text{ e } b \geq 0$$

É importante ressaltar que na forma padrão a função objetivo linear deve ser **minimizada**; as restrições do problema são definidas por um **sistema de equações lineares**; as condições de **não negatividade** de todas as variáveis de decisão complementam as restrições do problema. Para empregar o algoritmo Simplex, o sistema na forma padrão é particionado em parte Básica e Não Básica tal que $x^T = [x_N^T | x_B^T]$; $c^T = [c_N^T | c_B^T]$; $A = [N|B]$, com I_B e I_N os índices das posições do particionamento Básico e Não Básico, respectivamente.
Na primeira iteração do algoritmo Simplex, o particionamento será sempre formado por $I_N = \{1, 2, ...n'\}$ e $I_B = \{n' + 1, ..., n\}$ com B = I (matriz identidade de dimensão $m \times m$).
A cada nova iteração, cálculos como
- solução corrente obtida por $B \cdot x_B = b$ ou seja $x_B = B^{-1} \cdot b$ e $x_N = \overrightarrow{0}$. Variáveis cujo índice pertence a I_N são sempre nulas, está é uma condição para garantir que a solução corrente seja um vértice da região viável.
- custo relativo c'_i, $i \in I_N$ obtido por $c'_i = c_i - \lambda^T \cdot a_i$ tal que $\lambda = B^{T^{-1}} \cdot c_B$ e a_i é a i-ésima coluna da matriz A. Verificação da regra de otimalidade: $c'_i > 0 \ \forall i \in I_N$. Se a regra de otimalidade foi atingida, a solução corrente é ótima. Caso contrário deve-se realizar a troca da base para obter um novo vértice. Identificar o índice $k \in I_N$ com menor custo relativo negativo.
- direção simplex y tal que $y = B^{-1} \cdot a_k$, sendo a_k a k-ésima coluna da matriz A.
- tamanho de passo ϵ tal que $\epsilon_z = x_z / y_z, z \in I_B$. Verificação da regra de Parada: Não Existe $\epsilon_j > 0$ para algum $j \in I_B$? Se esta regra de parada foi atingida, conclui-se que não há solução ótima. Caso contrário, a variável básica com o menor $\epsilon_j > 0$ deixará a base. Identificar o índice $s \in I_B$ tal que $\epsilon_s = min\{\epsilon_j > 0, \forall j \in I_B\}$.
Os novos contadores serão atualizados, considerando tal mudança e nova iteração será iniciada com $I_N = \{1, 2, ..., n'\} - \{k\} \cup \{s\}$ e $I_B = \{n' + 1, n' + 2, ..., m\} - \{s\} \cup \{k\}$

Formulação do Modelo

O aplicativo "Simplex Formas"([3]), é uma ferramenta essencial para a estruturação de modelos de programação linear. Ele permite que os usuários definam a função objetivo e as restrições do problema, retornando a forma padrão, os elementos matriciais envolvidos e o particionamento inicial. Este aplicativo é especialmente útil para estudantes que estão começando a aprender sobre programação linear, pois facilita a visualização e compreensão dos componentes fundamentais do modelo. A interface intuitiva permite a entrada de dados de forma simples e direta, auxiliando na construção das matrizes e vetores necessários para a aplicação do método Simplex (Figura 1).

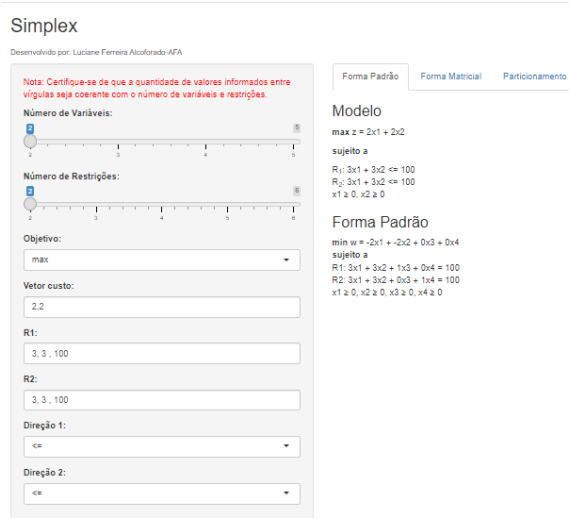


Figura 1: Aplicativo Simplex Formas. Fonte: [3], 2024.

Algoritmo Simplex

O aplicativo "Simplex Entra e Sai da Base"((Figura 2), permite ao usuário informar o número de variáveis e restrições, os coeficientes da função objetivo e restrições, as direções das restrições e seus limites. As entradas são exibidas e editáveis dinamicamente, sem a necessidade de um botão 'Submeter'. Além disso, o aplicativo auxilia nas iterações do algoritmo Simplex, fornecendo de modo automático o particionamento da primeira iteração. A partir da análise dos custos relativos c'_i e dos tamanhos de passo ϵ_j, o usuário pode realizar a mudança de base, alterando o I_B de entrada. Caso tenha dúvida, pode conferir na aba 'Próxima Iteração' a mudança que deverá ser feita.

Figura 2: Aplicativo Entra e Sai da Base. Fonte: [2], 2024.

3. RESULTADOS E CONCLUSÕES

Para ilustrar a eficácia dos aplicativos desenvolvidos face ao objetivo proposto, vamos considerar um exemplo prático de um problema de programação linear, que também pode ser resolvido pelo método gráfico conforme explicado em [1] e que agora será detalhado com o método do algoritmo:

$$\textbf{Maximizar } Z = 3x_1 + 2x_2$$
Sujeito a:
$$x_1 + x_2 \leq 4$$
$$2x_1 + x_2 \leq 5$$
$$x_1, x_2 \geq 0$$

Passo 1: **Formulação do Modelo:** O usuário inicia o aplicativo "Simplex Formas", inserindo a função objetivo $Z = 3x_1 + 2x_2$ e as restrições do problema. O aplicativo apresenta a forma padrão e os elementos matriciais envolvidos, facilitando a visualização da estrutura do problema.
Passo 2: **Inserção dos Dados:** No aplicativo "Simplex Entra e Sai da Base", o aluno informa o número de variáveis (2) e restrições (2), inserindo os coeficientes da função objetivo considerando o problema na forma padrão de minimização (-3 e -2), os coeficientes das restrições (1, 1; 2, 1), as direções das restrições ($\leq$) e os limites das restrições (4 e 5). As entradas são exibidas e editáveis dinamicamente.
Passo 3: **Primeira Iteração:** O aplicativo auxilia na montagem da primeira iteração do método Simplex, fornecendo o particionamento inicial. O aluno pode analisar os custos relativos c_i' e os tamanhos de passo ϵ_j, e realizar a mudança de base alterando o índice básico (I_B) de entrada.
Passo 4: **Iterações Subsequentes:** Caso o aluno tenha dúvidas sobre a próxima iteração, ele pode conferir na aba 'Próxima Iteração' a mudança que deverá ser feita. O aplicativo continua a auxiliar nas iterações até que a solução ótima seja encontrada. Se o modelo inicial era de maximização, a solução ótima de z^* deverá ter ser sinal invertido, pois o problema resolvido é sempre de minimização (forma padrão considerada).
Os aplicativos foram desenhados para simplificar de modo significativo a manipulação de matrizes e a execução do algoritmo Simplex, permitindo que os alunos se concentrem na compreensão dos conceitos fundamentais. Com a capacidade de lidar com uma ampla variedade de problemas de programação linear, permitindo variar o número de variáveis e restrições, podendo explorar diferentes cenários e aprofundar sua compreensão do método Simplex. A utilização da linguagem R e dos aplicativos Shiny ilustram a possibilidade de uma abordagem inovadora para o ensino de programação linear, facilitando a aprendizagem de conceitos complexos e promovendo uma melhor assimilação dos conteúdos.
Assim, ao automatizar a construção e manipulação de matrizes, os aplicativos não apenas economizam tempo, mas também permitem que os alunos se concentrem na compreensão dos conceitos fundamentais do método Simplex, promovendo uma melhor assimilação dos conteúdos e uma maior motivação para explorar diferentes problemas de programação linear.

AGRADECIMENTOS

À Divisão de Ensino da Academia da Força Aérea pelo apoio ao Projeto de Pesquisa Portaria AFA n. 87/SPPC.

Referências

[1] L. F. Alcoforado. Programação linear no plano: uma proposta usando ggplot2. In M. J. G. Villamayor, editor, *X Xornada de Usuarios de R en Galicia*, Santiago de Compostela, outubro 2023.

[2] L. F. Alcoforado. Simplex entra e sai da base. `https://lucianefalcoforado.shinyapps.io/Simplex_Entra_Sai_Base/`, 2024. Aplicativo Shiny para o método Simplex. Acessado em: 12 de setembro de 2024.

[3] L. F. Alcoforado. Simplex formas. `https://lucianefalcoforado.shinyapps.io/Simplex_Formas/`, 2024. Aplicativo Shiny para o método Simplex. Acessado em: 12 de setembro de 2024.

[4] M. N. Arenales, V. A. Armentano, R. Morabito, and H. H. Yanasse. *Pesquisa Operacional*. Elsevier, Rio de Janeiro, 1 edition, 2011. Capítulo sobre o Método Simplex, incluindo a formulação original de Dantzig e desenvolvimentos subsequentes.

[5] P. M. Barros, C. M. Araújo, and J. A. Fernandes. Raciocínios de estudantes do ensino superior na resolução de tarefas sobre matrizes. In J. A. Fernandes, M. H. Martinho, J. Tinoco, and F. Viseu, editors, *Atas do XXIV Seminário de Investigação em Educação Matemática*, Braga, 2013. ESTiG - Instituto Politécnico de Bragança, Centro de Matemática – Universidade do Minho, CIEd – Universidade do Minho, Centro de Investigação em Educação da Universidade do Minho.

[6] A. C. Steinhorst. O processo de construção dos conceitos de matrizes, determinantes e sistemas lineares no ensino médio, utilizando a planilha como recurso: um estudo comparativo. Dissertação de mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2011. Orientador: Prof. Dr. Lorí Viali.

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

O papel de R en Investigación Mariña. Da xenónmica a estudo global dos océanos

Isabel Fuentes-Santos[1]

[1]Instituto de Investigacións Mariñas (IIM-CSIC), 36208, Vigo, España

RESUMO

O Instituto de Investigacións Mariñas (IIM-CSIC) é un dos principais centros de investigación mariña da Península Ibérica. Mirando cara ao Atlántico dende Vigo, o IIM-CSIC enfróntase ao reto de avanzar no coñecemento sobre o océano, dende as súas características físicas ata os alimentos que tomamos del e os procesos que interconectan todo o sistema. As liñas de investigación do IIM van dende a secuenciación de ADN ou estudos inmunolóxicos de especies mariñas, ata a análise e modelización das características tanto físicas como bioxeoquímicas de océanos e áreas costeiras. En cada un destes ámbitos empréganse distintas metodoloxias de análise e inferencia estatística e, por tanto, distintos paquetes de `R`.

Neste traballo facemos unha revisión dos paquetes de `R` que se usan habitualmente no IIM. Os paquetes `mgcv`, para o axuste de modelos GAM e GAMM, e `ggplot2`, para representación gráfica, son ferramentas comúns nas distintas áreas de investigación do centro. En ecoloxía mariña úsase inferencia bayesiana, coa axuda de paquetes como `INLA`, `SIBER`, `nicheROBER`. En xenómica e biotecnoloxía combínase o uso de software específico con paquetes como `Bioconductor`, `adegenet`, `qiime2R` e `phylosec`. Nembargantes, en oceanografía, con gran cantidade de datos espaciotemporais e direccionais, o uso de `R` é residual, polo que tamén exploramos a contribución potencial de técnicas estatíticas desenvolvidas nos últimos anos e implementadas en `R` a esta área de investigación.

Palabras e frases chave: Oceanografía, ecoloxía pesqueira acuicultura, xenómica, bioinformática, análise estatistica.

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

PROCESSO DE OTIMIZAÇÃO DE UMA CARTEIRA DE ATIVOS UTILIZANDO A LINGUAGEM R

Ariel Levy[1], Marcus Antonio Cardoso Ramalho[1] e Eduardo Camilo da SIlva[1]

[1] Universidade Federal Fluminense

RESUMO

Uma demanda recorrente dos investidores é a maximização dos retornos e a minimização dos riscos. Para tal objetivo, este estudo apresenta o processo de otimização de carteiras utilizando o índice Sharpe. Ao aplicar as técnicas de otimização por meio de pacotes do R, considera-se as restrições e objetivos específicos. Portanto, consiste num guia prático para a construção de portfólios otimizados usando ferramentas de código aberto. Concluiu-se que dada a simplicidade do processo este constitui-se num ponto de partida para análises mais elaboradas em finanças quantitativas.

Palavras-chave: Otimização de carteiras, Índice de Sharpe, Linguagem R, Risco e Retorno, Fronteira Eficiente.

1. INTRODUÇÃO

O objetivo primordial de qualquer investimento é a obtenção de lucro. No contexto de investimentos em renda variável, a geração de lucro, ou prejuízo, é determinada pelo valor investido e pelas flutuações nos preços dos ativos. A meta principal é maximizar o lucro dentro do capital investido, o que é usualmente mensurado através de retornos percentuais calculados sobre as variações de preço dentro de um horizonte temporal definido (e.g., diário, mensal, anual). Entretanto, essa busca por retornos deve ser ponderada pelo risco inerente a cada investimento. Logo, o objetivo ideal consiste em maximizar o lucro e minimizar o risco.

Este trabalho apresenta uma análise de otimização de um portfólio de ações utilizando o software R. O foco é demonstrar as ferramentas e técnicas de análise, utilizando um portfólio genérico como exemplo didático. Na prática, a seleção de ativos para compor um portfólio é um processo complexo que exige uma análise criteriosa de indicadores fundamentalistas e o estudo do histórico financeiro das empresas.

A próxima seção aborda o arcabouço teórico da moderna teoria de portfólios e o índice de Sharpe, ferramentas essenciais para a otimização de carteiras, buscando o

ponto ótimo entre risco e retorno. Em seguida, o capítulo demonstra como utilizar o software R e suas bibliotecas para construir carteiras otimizadas considerando diferentes restrições e objetivos. Adicionalmente, discute-se as limitações do índice de Sharpe e exploram-se outras métricas de risco relevantes para a otimização de carteiras.

2. Teoria de Carteiras

A moderna teoria de portfólios revolucionou a maneira como os investimentos são geridos. A ideia central reside no princípio fundamental de que investidores buscam maximizar retornos e minimizar riscos, porém, raramente estão dispostos a assumir riscos sem a expectativa de retornos maiores. Markowitz demonstrou que a combinação de ativos com diferentes perfis de risco e retorno permite a construção de uma carteira otimizada que maximiza o retorno para um dado nível de risco ou minimiza o risco para um determinado nível de retorno (Markowitz, 1952).

No contexto de um portfólio, o retorno é calculado como a média ponderada dos retornos individuais dos ativos, considerando seus respectivos pesos na carteira. Já o risco de um portfólio não é simplesmente a média ponderada dos riscos individuais (desvio padrão) dos ativos. Entra em cena um novo componente: a covariância.

A covariância mede como o retorno de um ativo se movimenta em relação ao movimento do retorno de outro ativo. Em outras palavras, a covariância captura a relação entre os retornos de diferentes ativos. Ela é comumente expressa pelo produto do coeficiente de correlação de Pearson, que varia de -1 a +1, e os desvios padrões dos dois ativos.

O número de termos da equação da variância cresce exponencialmente com o número de ativos na carteira. À medida que adicionamos mais ativos a um portfólio, o impacto dos termos de covariância na variância total da carteira aumenta consideravelmente, superando o impacto dos termos de variância individual dos ativos. Este fenômeno é a base da diversificação.

A diversificação, intuitivamente comparada a "não colocar todos os ovos em uma cesta", permite reduzir o risco de um portfólio sem necessariamente comprometer o retorno. Isso ocorre porque a eliminação parcial ou total dos riscos não-sistemáticos, aqueles específicos de cada ativo, torna o risco da carteira mais próximo do risco sistemático, aquele inerente ao mercado como um todo e não diversificável.

A fronteira eficiente representa graficamente o conjunto de carteiras que oferecem o máximo retorno esperado para cada nível de risco, ou o mínimo risco para cada nível de retorno esperado. A escolha da carteira ideal dentro da fronteira eficiente depende do perfil de apetite a risco de cada investidor. Investidores mais arrojados tendem a optar por carteiras com maior risco e retorno, enquanto investidores mais conservadores preferem carteiras com menor risco e retorno.

O índice Sharpe é uma métrica amplamente utilizada para avaliar o desempenho de investimentos, comparando o retorno de uma carteira com o risco assumido. Ele mede

o retorno adicional (prêmio de risco) obtido por unidade de risco em relação a um investimento livre de risco. O índice Sharpe permite comparar diferentes carteiras e identificar aquelas que oferecem o melhor retorno ajustado ao risco (Sharpe, 1964).

3. Otimização de carteiras com R

A otimização de carteiras de investimentos constitui prática comum em finanças quantitativas, impulsionada pela abundância de dados e desenvolvimento de ferramentas computacionais. O R (R Core Team, 2024) oferece pacotes para análise financeira, tornando-se instrumento relevante nesse processo.

A primeira etapa para otimizar uma carteira de investimentos em R é a obtenção dos dados históricos dos ativos que a compõem. Para isso, utiliza-se o pacote *quantmod* (Ryan & Ulrich, 2023), que permite a importação direta dos dados de plataformas financeiras online, como o *Yahoo Finance* (Yahoo, 2024). Em seguida, calcula-se os retornos logarítmicos dos ativos utilizando a função *Return.calculate* do pacote *PerformanceAnalytics* (Peterson et al., 2020).

O próximo passo é a otimização da carteira. Primeiramente adicionam-se restrições e objetivos à otimização utilizando a função portfolio.spec do pacote *PortfolioAnalytics* (Peterson et al., 2024). Restrições são limites impostos à composição da carteira, como a soma dos pesos dos ativos ser igual a 1 ou a imposição de um peso máximo para cada ativo. Objetivos, por sua vez, definem as metas da otimização, como maximizar o retorno esperado, minimizar o risco ou o índice Sharpe.

Após definir as restrições e os objetivos, a otimização é realizada utilizando a função *optimize.portfolio*. O argumento *optimize_method* desta função define o algoritmo de otimização a ser utilizado, enquanto *maxSR = TRUE* indica que o objetivo é maximizar o índice Sharpe.

Com os pesos otimizados em mãos, podemos calcular o retorno da carteira otimizada e compará-la com o *benchmark* através de um gráfico de linha utilizando a função *chart.TimeSeries* do pacote *PerformanceAnalytics*.

Um guia mais detalhado sobre a otimização de carteiras utilizando R, pode ser encontrado no livro A Inteligência Artificial nas Ciências de Dados (Alcoforado et al., 2024, Capítulo 1).

3. Conclusão

Apresentamos a otimização de carteiras de investimentos utilizando o índice Sharpe como critério de decisão e a linguagem de programação R como ferramenta de aplicação. Descrevemos o processo de importação de dados, cálculo de retornos, definição de restrições e objetivos, otimização da carteira e análise de resultados utilizando os pacotes *quantmod* (Ryan & Ulrich, 2023), *tidyverse* (Wickham et al., 2019), *PerformanceAnalytics* (Peterson et al., 2020) e *PortfolioAnalytics* (Peterson et al., 2024).

Apesar da relevância do tema e da capacidade da linguagem R em lidar com problemas complexos de otimização de carteiras, algumas limitações devem ser

reconhecidas. Não foram abordados critérios de seleção para ativos da carteira e nem mesmo os períodos indicados. Portanto, a utilização do processo indicado não constitui recomendação de investimento.

Ademais, o estudo focou no processo metodológico, utilizando o índice Sharpe como principal métrica de risco-retorno e sem explorar outros modelos e indicadores relevantes em finanças quantitativas. Pesquisas futuras podem expandir a análise, incorporando temas como o modelo de precificação de ativos de capital (CAPM), o modelo de precificação por arbitragem (APT), o índice de Treynor, além de considerar diferentes métodos de otimização, restrições, objetivos e análises de sensibilidade.

Referências:

Alcoforado, L. F., Santos, J. P. M. D., Levy, A., Longo, O. C., Universidade Federal Fluminense, Linares, J. L., & Kubrusly, J. Q. (Orgs.). (2024). A Inteligência Artificial nas Ciências de Dados. Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos.

Markowitz, H. (1952). Portfolio Selection. The Journal of Finance, 7(1), 77. https://doi.org/10.2307/2975974

Peterson, B. G., Carl, P., Bennett, R., Boudt, K., Zhao, X., Martin, R. D., Yollin, G., Varon, H., Feng, X., & Kang, Y. (2024). PortfolioAnalytics: Portfolio Analysis, Including Numerical Methods for Optimization of Portfolios (Versão 2.0.0) [Software]. https://cran.r-project.org/web/packages/PortfolioAnalytics/index.html

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., Cornilly, D., Hung, E., Lestel, M., Balkissoon, K., Wuertz, D., Christidis, A. A., Martin, R. D., Zhou, Z. "Zenith", & Shea, J. M. (2020). PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis (Versão 2.0.4) [Software]. https://CRAN.R-project.org/package=PerformanceAnalytics

R Core Team. (2024). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Ryan, J. A., & Ulrich, J. M. (2023). quantmod: Quantitative financial modelling framework. https://CRAN.R-project.org/package=quantmod

Sharpe, W. F. (1964). CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK*. The Journal of Finance, 19(3), 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Yahoo. (2024). Yahoo Finance. https://finance.yahoo.com/

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

R para a elaboración e visualización da estatística de vivendas familiares principais ocupados en Galicia

Esther López Vizcaíno[1], Isabel del Río Viqueira[1] e Solmary Silveira Calviño[1]

[1] Instituto Galego de Estatística (IGE)

RESUMO

O obxectivo deste traballo é describir en que medida o software R nos axudou na elaboración e visualización da estatística de vivendas familiares principais ocupadas en Galicia. Describiremos como se levou a cabo a elaboración do directorio destas vivendas e as ferramentas de visualización empregadas para contraste e análise estatística.

Palabras e frases chave: R,shiny,

1. INTRODUCIÓN

O coñecemento do número e as características das vivendas familiares en Galicia é fundamental para levar a cabo políticas públicas que teñan como destino os inmobles/vivendas nas que habitan os residentes en Galicia.
A fonte de información máis importante que actualmente proporciona información sobre a poboación, o Padrón Municipal de Habitantes (PMH), non ten uns identificadores únicos para as vivendas. Por outra banda, os ficheiros do Catastro Inmobiliario que proporciona a Dirección General de Catastro do Ministerio de Hacienda, conteñen a Referencia Catastral (RC) como identificador único de vivenda, pero non teñen información sobre se as vivendas están ocupadas, son secundarias ou baleiras. No Catastro tamén se dispón, en principio, da ubicación xeográfica precisa (coordenadas) da vivenda. Ademais, a información catastral debería estar bastante actualizada e libre de erros polas repercusións tributarias.
O principal inconveniente de empregar o Catastro é que a súa información non está enlazada coa poboación residente, non se sabe a priori a relación entre as persoas do Padrón e as vivendas/inmobles de Catastro, é dicir, non se sabe quen vive en cada referencia catastral. Outro problema de Catastro é que nalgúns casos falta a división horizontal que fai que as vivendas dun edificio compartan unha única RC.

2. OBXECTIVO

O obxectivo deste traballo é identificar as vivendas familiares ocupadas en Galicia e asignarlles un identificador único, que prevaleza no tempo. Esto permitirá ofrecer estatísticas sobre o número e as características das vivendas por parroquias, por exemplo coñecer a distribución por tipoloxía, superficie, etc.

3. FONTES DE INFORMACIÓN

Ficheiros do Catastro inmobiliario

Os ficheiros de Catastro inclúen varios tipos de rexistros. Os utilizados neste traballo son os seguintes:

- Rexistro de fincas. Cada rexistro fai referencia a unha propiedade ou parcela catastral, que é a porción de terreo delimitada no que están os inmobles e construcións asociadas ao mesmo. Inclúe a RC da parcela a catorce posicións. Este tipo de rexistro contén outro tipo de información moi importante e valiosa que xa se mencionou: Coordenadas xeográficas.
- Rexistro de construcións. Identifica cada un dos locais existentes na propiedade coa súa descrición física: superficie, antigüidade, tipoloxía, destino.
- Rexistro de bens inmobles. Identifica, a través da RC a vinte posicións, cada un dos inmobles dentro dunha parcela catastral. Cada ben inmoble (concepto legal) inclúe unha ou varios construcións (concepto físico). Poderíase pensar, a priori, que cada ben inmoble cunha tipoloxía de vivenda é unha vivenda en si, pero non sempre ocorre na realidade. Por exemplo, hai unha serie de bloques de vivendas que aparecen como un único ben inmoble en Catastro, pero no que hai máis dunha vivenda, é o que se coñece como falta de división horizontal.
- Rexistro de titularidade. Inclúe datos de identificación do inmoble, RC a vinte díxitos, xunto cos datos identificativos dos seus correspondentes titulares catastrais. Tamén inclúe datos do dereito do propietario sobre a propiedade.

Ficheiro do PMH

É o rexistro administrativo no que constan os veciños dun concello. Os seus datos constitúen proba da residencia no concello e do domicilio habitual no mesmo. A lexislación española sobre réxime local establece as normas para a formación do Padrón municipal, que corresponde aos concellos, e da obtención das cifras de poboación a partir da revisión do mesmo no 1 de xaneiro de cada ano, unha vez levada a cabo polo INE a coordinación dos padróns municipais. Este ficheiro contén a información do nome, apelidos, DNI ou identificador equivalente, enderezo e idade da persoa residente.

Rueiro do Censo Electoral

O rueiro contén toda a información que identifica plenamente as vías e tramos de vía que pertencen a cada sección censual. Trátase dun conxunto de catro ficheiros: ficheiro de vías, ficheiro de pseudovías, ficheiro de tramos de vías e ficheiro de unidades poboacionais. Os ficheiros, que son independentes para cada provincia, son os que o INE utiliza para fins do Censo Electoral.

Base de datos sociodemográfica

Nesta base de datos, elaborada polo IGE, están todas as persoas que nalgún momento tiveron algunha relación con Galicia, constatada mediante os rexistros administrativos dos que se dispón no IGE: PMH, Afiliados á Seguridade Social, Pensionistas contributivos da Seguridade Social, ... Esta Base de datos sociodemográfica é unha fusión de rexistros administrativos coa finalidade de ter un sistema de información que conteña datos socioeconómicos da poboación que nalgún momento tivo relación con Galicia. Disponse do lugar de residencia da persoa e características socioeconómicas como a idade, o ano de nacemento, se está afiliada á Seguridade Social en alta, se cobra unha pensión contributiva, etc.. Esta base de datos ten tamén as coordenadas xeográficas das persoas, que será o que empreguemos neste traballo.

Rexistro dos depósitos de fianza de arrendamentos

En virtude do Acordo entre o IGE e o Instituto Galego da Vivenda e o Solo (IGVS)

para intercambiar información sobre fianzas de alugueiros firmado en maio de 2024 o IGVS proporciónalle ao IGE un rexistro onde constan as fianzas depositadas polos arrendadores dos bens inmobles en Galicia, co cal contén información daquelas vivendas que están alugadas, identificadas mediante a súa RC. Nos contratos de arrendamento relativos a vivendas e predios urbanos será obrigatoria a esixencia e prestación de fianza en metálico. Teñen a obriga de depositar esta fianza os arrendadores de vivendas e predios urbanos ante o IGVS.

3. PROCEDEMENTO SEGUIDO PARA O CRUCE DE TODAS AS BASES DE DATOS

O proceso para a identificación e caracterización das vivendas familiares segue catro pasos:

1. En primeiro lugar definirase un directorio de vivendas a partir dos datos do Catastro Inmobiliario. O Catastro non identifica nin contabiliza vivendas, se non que identifica bens inmobles con uso residencial ou construcións con destino residencial. O ben inmoble é unha unidade xurídica, que o Catastro divide en construcións, en función das peculiaridades destas para definir con precisión as características dos inmobles e poder asignarlle a correspondente valoración catastral. A vivenda, sen embargo, é unha unidade física, tal e como se quere identificar neste traballo. Para chegar a este concepto a partir de vivenda e dende os bens inmobles foi necesario facer determinados procesos previos, en parte baseados no traballo de Enrique et al. [1]. A dificultade que presenta traballar con estas bases de datos é a dimensión das mesmas, por esta razón neste traballo empregouse o paquete de R **dbplyr** [5].

2. A continuación farase unha asignación entre as vías do Catastro e as vías do Rueiro do Censo Electoral. É necesario determinar unha relación das vías de Catastro co Rueiro do Censo Electoral (utilizado no PMH) para poder identificar aquelas vivendas que son residencia habitual da poboación de Galicia. Cada persoa pode ter varias vivendas en propiedade (Catastro) -resida ou non nelas- e ao contrario, unha persoa pode residir (PMH) nunha vivenda que pode ter ou non en propiedade (pode ser alugada ou cedida); polo tanto, a relación entre os ficheiros de Catastro e de PMH deberá efectuarse empregando o enderezo das vivendas. Neste procedemento fanse moitas comparacións de textos e emprégase principalmente o paquete de R **text2vec** [3]. Tamén se empregou ao paquete de R **sf** [6] para a xeración de envolventes convexas de vivendas co obxectivo de contrastar que as asignacións anteriores estivesen correctas.

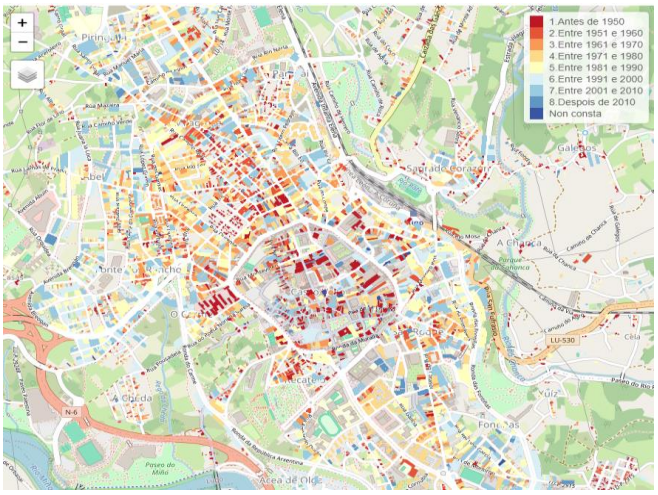
3. Neste punto determinaranse as vivendas familiares principais no PMH. O PMH indica, para cada persoa, cal é o seu fogar padronal. Este fogar vén dado polas persoas empadronadas nun mesmo enderezo, ou, cando non está suficientemente determinado, polas persoas inscritas na mesma folla padronal. Neste procedemento empregaremos principalmente os paquetes **dplyr**, **purrr**, [5] e **fuzzyjoin** [8].

4. Por último, asignaranse as RC ás vivendas principais do PMH. Neste punto está dispoñible o directorio de vivendas coas súas características e a súa titularidade de Catastro. Ademais, disponse da relación entre os códigos de vía de Catastro e os códigos de vías do PMH, obtidos nun procedemento anterior. Tendo en conta que o 78% das persoas galegas son propietarias da vivenda onde viven (IGE, 2020), cruzaremos o ficheiro da residencia das persoas do PMH (coas vivendas principais identificadas no paso anterior) co directorio de vivendas de Catastro, para asignarlle ás persoas a RC da vivenda onde residen. Neste procedemento empregaremos principalmente os paquetes do entorno **tidyverse** [5], **stringdist** [4], **sf** [6]

5. FERRAMENTAS DE VISUALIZACIÓN

Unha vez que se dispón do directorio de vivendas familiares principais, é importante ter ferramentas de visualización que axuden, por un lado, a contrastar visualmente todos os traballos anteriores e, polo outro, a visualizar información agregada que nos achegue conclusións sobre o espazo habitado de Galicia.
Neste punto botaremos man das librerías **shinydashboard** [9] e **tmap**[10]**.** Na Figura 1 presentamos un exemplo da visualización empregada.

Figura 1. Vivendas familiares principais ocupadas no concello de Lugo segundo o ano de construción.



Referencias

[1] Enrique, I., Valverde, J, Ramirez, A., Ojeda, S. (2020) Identificación de las viviendas y sus características en la información del Catastro. El caso de Andalucía. Revista Catastro, 99.

[2] IGE (2020) Enquisa estrutural a fogares. Vivendas familiares. Características e medio: https://www.ige.gal/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0304005.

[3] Selivanov D., Bickel, M., Wang, O. (2022) text2vec: Modern Text Mining Framework for R. R package version 0.6.3. https://CRAN.R-project.org/package=text2vec.

[4] Van der Loo, M. (2014) The stringdist package for approximate string matching. The R Journal, 6, 111-122. https://CRAN.R-project.org/package=stringdist.

[5] Wickham H, Girlich M, Ruiz E (2023) dbplyr: A 'dplyr' Back End for Databases. R package version 2.3.2, https://CRAN.R-project.org/package=dbplyr.

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

[7] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686.

[8] Robinson D (2020). _fuzzyjoin: Join Tables Together on Inexact Matching_. R package version 0.1.6, https://CRAN.R-project.org/package=fuzzyjoin

[9] Chang W, Borges Ribeiro B (2021). _shinydashboard: Create Dashboards with 'Shiny'_. R package version 0.7.2, <https://CRAN.R-project.org/package=shinydashboard>

[10] Tennekes M (2018). "tmap: Thematic Maps in R." _Journal of Statistical Software_, *84*(6), 1-39. doi:10.18637/jss.v084.i06 <https://doi.org/10.18637/jss.v084.i06>

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Comparison of cumulative incidence curves with mçultiple causes of death

Nora M. Villanueva[1], Marta Sestelo[2,1], Luís Meira-Machado[3] and Javier Roca-Padiñas[2,1]

[1] Dep. Statistics and O.R., & SiDOR group, University of Vigo, Vigo
[2] CITMAga, Santiago de Compostela, 15782, Spain.
[3] Centre of Mathematics & Department of Mathematics, University of Minho, Portugal.

RESUMO

The comparison of multiple populations using different curves is a topic that arises frequently in many areas. This is particularly important in the medical field where one often wants to compare multiple survival curves [2, 3]. Though this can be performed using parametric models through the comparison of the resulting model parameters, nonparametric methods are usually used for this purpose. However, the proposed methods cannot be applied to competing risk survival data and there are few methods to compare groups or curves of interest in the presence of competing risks. Most of the literature is focused on the comparison of the cumulative incidence functions for a particular type of failure among different groups. However, the comparison of the cumulative incidence functions among each other is also useful since it can reveal whether two or more of these functions are equal or whether one is "more serious" than the other. With this in mind, we propose an approach that allows determining clusters of cumulative incidence functions with an automatic selection of their number [1].

Palabras e frases chave:: Multiple curves; Cumulative Incidence Functions; Clustering; Competing Risk data; Causes of Death.

Referencias

[1] Marta Sestelo, Luís Meira-Machado, Nora M. Villanueva, and Javier Roca-Pardiñas. A method for determining groups in cumulative incidence curves in competing risk data. *Biometrical Journal*, 66(4), 2024.

[2] Nora M. Villanueva, Marta Sestelo, and Luís Meira-Machado. A Method for Determining Groups in Multiple Survival Curves. *Statistics in Medicine*, 38(5):366–377, 2019.

[3] Nora M. Villanueva, Marta Sestelo, Luís Meira-Machado, and Javier Roca-Pardiñas. clustcurv: An R Package for Determining Groups in Multiple Curves. *The R Journal*, 13(1):164–183, 2021.

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Mandalas Matemáticas: Uma Releitura das Padronagens e Cores das Cerâmicas de Sargadelos

Luciane Ferreira Alcoforado[1], João Paulo Martins dos Santos[1] e Maria Cláudia de Jesus Machado [1]

[1]Academia da Força Aérea, Pirassununga, São Paulo, Brasil.

RESUMO

Este trabalho apresenta a criação de mandalas inspiradas nas cerâmicas de Sargadelos, combinando análise histórica, padrões geométricos e paletas de cores tradicionais com curvas matemáticas e transformações geométricas. O processo resulta em mandalas que reinterpretam a herança visual de Sargadelos, unindo história, matemática e computação.

Palavras chave: Mandalas. Padrões Geométricos. Cerâmicas de Sargadelos.

1. INTRODUÇÃO

Considerando a arte uma forma de comunicação, entende-se que ela constitui não apenas a expressão da beleza e da harmonia, mas também da expressão da identidade de um povo: sua história, seus costumes, suas crenças, seu habitat. A arte da cerâmica manifesta-se na cultura dos povos desde a mais remota antiguidade. Na Galícia destaca-se a cerâmica de Sargadelos, cuja pintura possui padrões geométricos tipicamente na cor azul cobalto em fundo branco. As cores verdes e ocre também estão presentes, escolha que se justifica pelo modo de cozimento: "Los colores azules, verdes y ocres son óxidos metálicos que cubren las piezas de Sargadelos en su mayor parte y resultan ser los colores más idóneos para cocer bajo cubierta a tan altas temperaturas como exige la porcelana." [7].
A história da produção da cerâmica de Sargadelos remonta ao século XIX impulsionada por Antonio Raimundo Ibáñez, Marques de Sargadelos, com a criação da primeira fábrica na Galícia. Ressalta-se que a matéria prima abundante na região para a elaboração da cerâmica era o caulim (argila branca). No entanto, no século XX, no pós-guerra espanhol, os intelectuais Isaac Díaz Pardo y Luis Seoane López iniciam um projeto de grande dimensão, que levaria à criação do complexo industrial do Laboratório de Formas em 1963. Esse projeto visou, sobretudo, revitalizar a produção de cerâmica e, por meio dela, resgatar a história da terra galega e sua identidade, assim como a memória histórica do lugar, que tinham sido desvalorizadas como consequência da repressão franquista. Inspiram-se em formas existentes na paisagem, no mundo rural, nos objetos herdados do passado, nas profissões, nas igrejas, nos monumentos. Todas estas formas refletem as características culturais locais, mas agregando traços de modernidade. Destaca-se que o tipo de produção da cerâmica envolve processos mecânicos e manuais. Portanto, tradição e renovação se fundem nas formas, no estilo desta cerâmica e na maneira como é produzida.

> "El Laboratorio de Formas era la agrupación de una serie de proyectos, algunos generados en la última etapa de la dictadura franquista, otros impulsados durante la transición democrática; los cuales se basaban en la idea de experimentación, la renovación de las producciones populares, la recuperación del patrimonio material e inmaterial, incorporando emblemas del pasado, de la civilización celtíbera que entroncaban con las costumbres y la memoria" [8]

As joias, bem como os amuletos são exemplos desta fusão. Acredita-se que os amuletos, na forma de pingentes, inspirados nas antigas lendas celtas, oferecem proteção contra as bruxas ou meigas, protagonistas do folclore galego, repleto de magia e mistério. Estas personagens femininas estão associadas com feitiços malignos em muitas histórias e, ao mesmo tempo, com habilidades mágicas, capazes de curar. Por conseguinte, os desenhos que servem de inspiração para a criação dos pingentes defensivos são uma fusão da história, geografia e cultura do povo galego com um tipo de arte com traços mais contemporâneos.
Neste contexto, considerando os padrões geométricos e cores típicas presentes nas tradicionais cerâmicas de Sargadelos, este trabalho se propõe a produzir um conjunto de mandalas com base na integração entre os elementos histórico-culturais da Galícia e os elementos Matemáticos e Computacionais.

2. METODOLOGIA

A metodologia básica para a construção das mandalas foi baseada no processo desenvolvido em [1], [2] e complementado em [3], [4]. Um elemento essencial é que os autores pressupõem que uma curva seja conhecida, em geral na forma paramétrica. No presente artigo, esse pressuposto é violado, ou seja, um padrão de cores e formas é apresentado e é necessário obter as figuras e cores elementares para o processo de construção. Dessa forma, devido ao viés de associação histórica, foi necessário um levantamento para estabelecimento das possibilidades de padrões e cores. Após, foi necessário estabelecer os padrões de cores aproximados baseados em inspeção visual das cores do modelo RGB, disponíveis no R básico (Ver [5]); as paletas de cores foram construídas de forma empírica para refletir a coloração observada nas amostras de figuras consultadas. A identificação das curvas que permitiram recriar, ao menos aproximadamente, os elementos geométricos da amostra, são objetos de análise da fase subsequente. Neste momento o processo de construção converge para o delineamentos de [2], o qual aplica as transformações geométricas tais como rotações, translações, reflexões e homotetias.
A paleta de cores é composta de três outras paletas de forma que é possível estabelecer proporções das cores adotadas. Os pormenores podem ser consultados nos códigos disponibilizados no GitHub. Deve ser notado, no entanto, que apesar do delineamento dos elementos construtivos, há fases críticas no desenvolvimento, sendo que a principal é o estabelecimento de uma ou mais curvas, de forma empírica, que irão possibilitar a criação.

3. RESULTADOS E CONCLUSÕES

O levantamento da paleta de cores foi realizado com base em amostras do Catálogo da Galeria Oficial de Sargadelos, conforme Figura 2. A amostra de símbolos utilizada para a construção das mandalas temáticas estão mostradas na Figura 1.

Figura 1: Amostra de padrões utilizados nas cerâmicas de Sargadelos.
Fonte: blog Art natura galicia, 2018.

Figura 2: Seleção da paleta de cores típica de Sargadelos
Fonte: Catálogo da Galeria Oficial de Sargadelos, 2024.

As mandalas inspiradas na coluna "Amostra"da Tabela 1 foram desenvolvidas a partir da equação paramétrica da circunferência $x = r \cdot cos(\theta)$ e $x = r \cdot sin(\theta)$. Para a amostra 1 utilizou-se de translações (`MandalaR::f_transxy`) intercaladas com rotações sucessivas (`MandalaR::f_rotacao`) e filtro (`dplyr::filter`) para pontos cuja distância fosse menor ou igual a 1 e finalizando com reduções (`MandalaR::f_factor`) entre 1% e 90%, gerando o conjunto final de pontos armazenados em um *dataframe*. Para obter o efeito de coloração foi utilizado uma função criada para esta finalidade denominada de `colorir_mandala` que tem como argumento outra função de criação de paleta de cores, o que possibilita o emprego de paletas criadas para cada caso. Especificamente para este estudo foi criada a função `gerar_paleta_sargadelos` que contém as três bases de cores de acordo com a Figura 2. Para a amostra 2 utilizou-se de circunferência com oito diferentes raios (r0=0.1, r1=0.3, r2=0.4, r3=0.5, r4=0.7, r5=0.8, r6=0.9, r7=1) e identificação dos setores circulares para aplicação do método de coloração que seguiu um conjunto de 4 regras baseadas nas distâncias entre os raios, assim uma coluna contendo a cor de cada ponto foi acrescentada ao dataframe final. Para obter o desenho final, foram utilizadas três camadas com as funções `ggplot2::geom_point`, `ggplot2::geom_segment` e `ggplot2::geom_curve`. Para a amostra 3, identificou-se que trata-se de um tipo de nó Celta, cuja curva paramétrica não foi encontrada, desse modo, utilizou-se de uma construção com círculos e segmentos de reta que procurasse expressar de forma aproximada o desenho da amostra. Os códigos em R destas construções específicas estão disponíveis em GitHub.

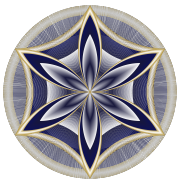
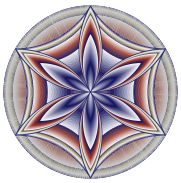
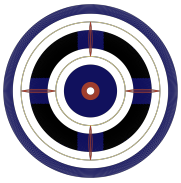
Amostra	Mandala 1	Mandala 2	Mandala 3
			
			
			

Tabela 1: Mandalas inspiradas nas cerâmicas de Sargadelos. Fonte: Autores, 2024.

As mandalas inspiradas na coluna "Amostra"da Tabela 2 foram desenvolvidas a partir das composições entre circunferências, da Lemniscata de Gerono, a função seno e a espiral de Euler, seguindo as referências apontadas e paleta de cores utilizada anteriormente. A primeira utiliza rotações da Lemniscata de Gerono composta com círculos concêntricos; a segunda utiliza circunferências concêntricas com sequências de raios agrupadas em torno de duas regiões distintas e rotações da função seno ($\sin(x)$) definido em intervalo $[a, b]$ específico; por fim, a terceira utiliza uma aproximação em série da Espiral de Euler com exclusão de pontos em duas regiões específicas. Os pormenores dos códigos e aproximações utilizadas podem ser encontrados no GitHub.

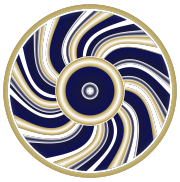
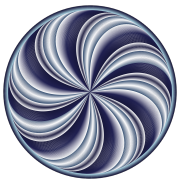
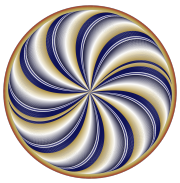
Amostra	Mandala 1	Mandala 2	Mandala 3
			
			
			

Tabela 2: Mandalas inspiradas nas cerâmicas de Sargadelos. Fonte: Autores, 2024.

Referências

[1] Alcoforado, L. F. (2022). Construindo Mandalas com R. In María José Ginzo Villamayor (Eds.), *IX Xornada de Usuarios de R en Galicia*. Santiago de Compostela, 20 out. 2022. Disponível em: https://www.r-users.gal/sites/default/files/libro2022.pdf. Acesso em: 12 set. 2024.

[2] Alcoforado, L. F., Santos, J. P. M., Lima, M. V. A., Firmiano, A., & López Linares, J. (2023). *Mandalas, curvas clássicas e visualização com R*. Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos. Disponível em: https://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/book/1017. Acesso em: 1 ago. 2023.

[3] Santos, J. P. M. (2024). Curvas e Cores em R: Movimentos Rígidos no Plano. In L. F. Alcoforado *et al.* (Eds.), *Aplicações em R: encurtando distâncias nas ciências* (pp. 123-145). Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos. Disponível em: http://www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/1249. Acesso em: 27 jul. 2024.

[4] Santos, J. P. M., & Alcoforado, L. F. (2023). Colorindo mandalas com R: explorando cores e gradientes em curvas planas. In María José Ginzo Villamayor (Eds.), *X Xornada de Usuarios de R en Galicia*. Santiago de Compostela, 18 out. 2023. Disponível em: https://www.r-users.gal/sites/default/files/libro2023.pdf.

[5] R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: https://www.R-project.org/.

[6] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. Disponível em: https://ggplot2.tidyverse.org.

[7] STEBAN GÓMEZ, Teresa. *Evolución y cambio de las formas cerámicas en Sargadelos: orígenes y características de la empresa cerámica sargadeliana*. Tese de doutorado. Universidad Complutense de Madrid. 1992. ISBN: 978-84-8466-191-7.

[8] REAL LÓPEZ, Inmaculada. *Sargadelos: patrimonio cultural, memorístico y turístico*. I Simposio anual de Patrimonio Natural y Cultural ICOMOS España. 21-23 de noviembre 2019. Madrid. DOI: `https://doi.org/10.4995/icomos2019.2020.11405`.

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

KEEP THE BALL ROLLING: CONTROL ESTADÍSTICO DE LA CALIDAD PARA LA INDUSTRIA 5.0

Salvador Naya[1], Javier Tarrío-Saavedra[1] y Miguel Flores[2]

[1]Grupo MODES, CITIC, Departamento de Matemáticas, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña.

[2] Departamento de Matemática. Grupo MODES. Escuela Politécnica Nacional. Quito. Ecuador.

RESUMO

En este trabajo queremos presentar ejemplos de aplicación con R usados en clase y que forman parte del libro "control estadístico de calidad en la Industria 5.0". El objetivo es exponer una serie de ejemplos de aplicación en el contexto actual de la denominada Industria 5.0. En el libro, todos los ejemplos presentados están resueltos usando el Software R, concretamente los paquetes qcr, ILS y r6qualityTools que incorporan técnicas novedosas de tipo no paramétrico tanto para datos multivariante como para datos funcionales, muy habituales en este nuevo contexto de la Industria 5.0, marco en que los datos son obtenidos en gran parte mediante sensores y la aplicación de técnicas IoT.

Palabras e frases chave: Control Estadístico de la Calidad, DoE, Gráficos de control, Análisis de Capacidad, Industria 5.0.

1. INTRODUCCIÓN

La Industria 5.0 representa un nuevo paradigma, que va más allá de la mera automatización y digitalización de procesos, centrándose en la colaboración entre humanos y máquinas para lograr una producción más flexible, personalizada y sostenible. En este contexto, el control estadístico de la calidad juega un papel fundamental para garantizar la excelencia en los procesos productivos y la satisfacción del cliente. Podemos resumir la Industria 5.0 como una ampliación de la 4.0 en la que se incorporan nuevas claves como: Colaboración humano-máquina, la Sostenibilidad y la Resiliencia. Los elementos clave son la automatización, la robotización, el BigData, los sistemas inteligentes, la virtualización, la IA, el aprendizaje automático y la Internet de las cosas (IoT).

Según la Unión Europea, este concepto "ofrece una visión de la industria que va más allá de la eficiencia y la productividad como únicos objetivos, y refuerza el papel y la contribución de la industria a la sociedad" y "sitúa el bienestar del trabajador en el centro del proceso de producción y utiliza las nuevas tecnologías para proporcionar prosperidad más allá del empleo y el crecimiento, respetando los límites de producción del planeta". Por tanto, la Industria 5.0 complementa el enfoque de la Industria 4.0

"poniendo específicamente la investigación y la innovación al servicio de la transición hacia una industria europea sostenible, centrada en el ser humano" [1,2].

Bajo este nuevo paradigma, el control estadístico de la calidad en la Industria 5.0 debe adaptarse a las nuevas realidades y desafíos que presenta este paradigma. Por un lado, adaptando los procesos al manejo de datos complejos, la interconexión de sistemas y el uso de sensores generan grandes volúmenes de datos complejos y heterogéneos. El control de calidad debe ser capaz de procesar y analizar estos datos de manera eficiente. Por otra parte, el análisis da datos en tiempo real, que precisa de la toma de decisiones en tiempo real es crucial en la Industria 5.0. Los métodos de control de calidad deben proporcionar información instantánea para permitir ajustes inmediatos en los procesos productivos.

Como novedad, se presentarán técnicas avanzadas para datos funcionales además de herramientas de la estadística no paramétrica para datos multivariantes, con., con aplicaciones a los procesos de la Industria 5.0. Es habitual que, este tipo de procesos (debido al uso de sensores), generen datos de tipo funcional. Por otra parte, es frecuente que los datos no siempre sigan distribuciones paramétricas conocidas, por lo que es necesario utilizar técnicas estadísticas avanzadas para manejar estos tipos de datos complejos [3].

2. PAQUETES DE R PARA EL CONTROL DE CALIDAD EN LA INDUSTRIA 5.0

Los paquetes qcr e ILS de R ofrecen herramientas versátiles y con gran capacidad para el control estadístico de la calidad en el contexto de la Industria 5.0. [4, 5, 6, 7].

El paquete qcr permite tanto aplicar herramientas básicas del control de la calidad como también gráficos de control avanzados. Entre las novedades está la opción de implementar gráficos de control para datos funcionales.

En relación al análisis de capacidad de procesos para cumplir especificaciones, el paquete *qcr* proporciona métodos para el cálculo de índices de capacidad para cumplir especificaciones y su representación gráfica, con el objeto de evaluar la capacidad de procesos complejos. Como novedad incluye técnicas robustas para el cálculo de índices de capacidad.

El paquete *ILS*, centrado inicialmente en los estudios interlaboratorio (Inter Laboratory Study), facilita el análisis de estudios de comparación entre laboratorios, esenciales para la calibración y validación de métodos en la Industria 5.0.

Como novedad la librería ILS incorpora la opción de usar métodos no paramétricos y ofrece técnicas estadísticas que no asumen distribuciones específicas, adaptándose a la naturaleza diversa de los datos en la Industria 5.0. Igualmente, proporciona aproximaciones para datos funcionales de los *estadísticos h y k de Mandel*. Por tanto, el paquete *ILS* presenta al usuario diversas alternativas robustas para la detección de laboratorios atípicos en los contextos del análisis multivariante y el Análisis de Datos Funcionales (FDA).

Por otra parte, el paquete r6qualityTools proporciona un conjunto completo de herramientas estadísticas para la gestión de la calidad, diseñado en torno al ciclo Definir, Medir, Analizar, Mejorar y Controlar (DMAIC), utilizado en la metodología Six Sigma. Es importante destacar que esta librería representa una actualización del paquete qualitytools, actualmente retirado del CRAN y archivado [8]. Este paquete tiene como novedad la programación orientada a objetos 'R6' para una mayor flexibilidad y rendimiento. Reemplaza los gráficos tradicionales con visualizaciones modernas e interactivas utilizando 'ggplot2' y 'plotly'. Desarrollado sobre los principios de 'tidyverse', simplifica la manipulación y visualización de datos, ofreciendo un enfoque intuitivo a la ciencia de la calidad. [9].

3. APLICACIONES PRÁCTICAS

Se presentarán distintas aplicaciones a problemas reales con los que han trabajado los autores y que se centran en el control de procesos en tiempo real. Se analizaron el empleo de gráficos de control del paquete *qcr* para monitorear variables críticas en líneas de producción automatizadas. Se presentará un estudio de la capacidad de un

proceso en fabricación aditiva, empleando técnicas de análisis funcional del paquete *qcr* para evaluar la calidad de piedra artificial.

Se presentará un ejemplo de validación de sensores inteligentes aplicando los métodos del paquete *ILS* para comparar y validar las mediciones de diferentes sensores en un entorno de fabricación inteligente.

Finalmente se presentarán ejemplos de aplicación del paquete *r6qualitytools* para el estudio de procesos siguiendo el modelo SixSigma. Concretamente, se presenta un caso de estudio en el que se analizan los datos de tránsito de varios buques por el Canal de Panamá Ampliado. Este ejemplo se usa en el libro para un estudio detallado de estadística descriptiva para finalmente presentar un modelo de regresión que facilita el estudio de la llamada *curva de aprendizaje* de los prácticos del Canal. También se emplea este paquete para el cálculo de *Gráficos de Control tipo Shewhart* para medidas individuales, obtenido a partir de datos sintéticos del tránsito de buques a través de la tercera esclusa del Canal de Panamá. Se usarán los gráficos obtenidos para monotorizar el proceso de tránsito.

Una segunda aplicación práctica es el estudio de control de calidad de cigüeñales de automóvil, en este caso usamos el paquete *SixSigma* para analizar los defectos superficiales y las grietas. Concretamente se empleará la función *ss.pMap* para la obtención del *mapa de procesos* correspondiente a la producción de cigüeñales. Este es otro ejemplo real de aplicación en datos procedentes de una empresa y que se pondrá de ejemplo para presentar las salidas gráficas de varios paquetes que presentan mapas de procesos y otras herramientas básicas del control de procesos industriales como los *diagramas de Ishikawa* o los *diagramas de Pareto*.

Otro conjunto de datos, son los de un estudio de dureza de un nuevo material que simula la piedra, que está disponible en el paquete *qcr*, en el *data.frame plates*. Con estos datos se analizarán diferentes gráficos de control y el estudio de la capacidad del proceso de producción.

Para el caso de datos funcionales, se hará uso de los datos sobre eficiencia energética del *data.frame*, *Consumo*. Con estos datos es posible aplicar gráficos de tipo funcional. Además, se incorporan nuevas funciones para la estimación de índices de capacidad no paramétricos.

En lo que respecta al control de calidad multivariante, se hará del paquete qcr, y del conjunto de datos *Contaminacion.txt*. Este conjunto de datos sintéticos contiene dos variables: por un lado el contenido en contaminantes, específicamente agua (ppm) y por otra, partículas sólidas (mg/l) en el queroseno usado como combustible para aviación. El contenido máximo de estos contaminantes esta regulado por normas internacionales; un alto contenido en agua o partículas solidas pueden ser fuente de fallos en los elementos de las aeronaves y, por tanto, de accidentes. Esta base de datos está basada en experimentos reales realizados en laboratorio.

Finalmente se dedica un capítulo al diseño de experimentos (DoE) y otro al estudio de la detección de anomalías a partir de datos complejos. Se emplean datos de experimentos realizados en laboratorio en los que se usa el paquete, *ILS*, para el estudio de la consistencia entre laboratorios, *diseños RyR* que se extienden también al estudio de consistencia de datos proveniente de sensores en los que se proponiendo en este caso el cálculo de aproximaciones FDA de los estadísticos *h y k de Mandel*, definidos para la detección de laboratorios atípicos.

4. CONCLUSIONES

El control estadístico de la calidad es esencial para el éxito de la Industria 5.0, permitiendo mantener altos estándares de calidad en entornos de producción altamente personalizados y colaborativos. Los paquetes qcr e ILS de R proporcionan herramientas avanzadas que se adaptan perfectamente a los desafíos de este nuevo paradigma industrial. Para profundizar en estos temas y explorar más aplicaciones prácticas, recomendamos nuestro próximo libro "Control de Calidad Avanzado para la Industria 5.0: Aplicaciones con R". Esta obra ofrece una guía completa sobre cómo implementar técnicas de control de calidad utilizando R, con un enfoque especial en

los paquetes qcr e ILS. El libro incluye numerosos ejemplos prácticos que demuestran cómo estas herramientas pueden aplicarse en escenarios reales de la Industria 5.0, desde el control de procesos automatizados hasta la validación de sistemas de inteligencia artificial en entornos de fabricación.

AGRADECIMIENTOS

Esta publicación/trabajo es parte de los proyectos de I+D+i PID2020-113578RB-I00 y PID2023-147127OB-I00 "FEDER/UE", financiados por MCIN/AEI/10.13039/501100011033/. También ha sido financiado por la Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2024/14) y por el CITIC como centro con acreditación en excelencia del Sistema universitario de Galicia y miembro de la Red CIGUS, y subvencionado por la Consellería de Educación, Ciencia, Universidades e Formación Profesional da Xunta de Galicia y cofinanciado a través del Fondo Europeo de Desenvolvemento Rexional (FEDER), Programa operativo FEDER Galicia 2021-27, (Ref. ED431G 2023/01).

Referencias

[1] Naya, S., Tarrío-Saavedra, J, Flores, M. (2004). Control estadístico de la calidad para la Industria 5.0. AulaMagna. McGrawHill.

[2] European Commission. (2021) "Industry 5.0", 2021. https://ec.europa.eu/info/research-and-innovation/research-area/industrial-research-and-innovation/industry-50_en.

[3] Bartaa, J., Kayserb I. (2023). Industry 5.0. Past, Present, and Near Future. Procedia Computer Science 219, 778–788.

[4] Flores, M. Fernandez-Casal, R. Naya, S. Tarrío-Saavedra,J. Statistical quality control with the qcr package, R Journal 13 (1) (2021) 194–217.

[5] Flores, M. Fernandez-Casal, R. Naya, S. Tarrío-Saavedra, J. Bossano, B. ILS: An r package for statistical analysis in interlaboratory studies, Chem metrics and Intelligent Laboratory Systems 181 (2018) 11–20.

[6] Flores, M. Fernandez-Casal, R. Naya, S. Tarrío-Saavedra,J. qcr: Quality Control Review, r package version 1.4 (2022).

[7] Flores, M. Fernandez-Casal, R. Naya, S. Tarrío-Saavedra,J. ILS: Interlaboratory Study, r package version 0.3 (2023).

[8] Roth, T. (2016). qualityTools: Statistics in Quality Science. R package version 1.55.

[9] Barahona, A., Encarnacion, F, Flores, M., Tarrío-Saavedra, J. and Naya, S. (2024). R6-Based Statistical Methods for Quality Science. Package 'r6qualitytools'.

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

Calculo de rutas de escape nun incendio forestal

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

[1]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga)
[2]Departamento de estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

RESUMO

Neste traballo presentase un algoritmo desenvolto no marco do proxecto Civil UAVs Initiative (CUI) da Xunta de Galicia e a empresa Avincis Aviation Spain SA. Este algoritmo ten por obxetivo calcular as rutas mais rápidas de evacuación que poderían utilizar os brigadistas que traballan nun incendio. Para isto o algoritmo ten en conta a pendente, as vías de transporte, as masas de auga e os modelos de combustible presentes da rexión do incendio, que poden estar en local ou obtidas a través dos servizos da IDEE. O algoritmo emprega multitude de librarías de R de estatística espacial como *sf*, *terra* e *tidyterra* e outras librarías coñecidas como a libraría *dplyr*. Tamén se utilizou docker para oseu paso a producción, o que facilita a súa execución en distintos equipos.

Palabras e frases chave: Ruta escape, incendio, grafo, docker, brigadas, datos espaciais.

1. INTRODUCIÓN

Na extinción dun incendio forestal un elemento de gran importancia é o uso de brigadas para o control do mesmo dende terra. Para iso é necesario manter unha boa comunicación e organización das mesmas para enfrentarse con seguridade ao incendio. Un perigo que afrontan as brigadas é que o incendio, na súa evolución chegue a cercalos. Esta situación pode darse porque o terreo situado na retaguardia das brigadas vaise secando polo efecto da enerxía que desprende o incendio, de maneira que aumenta a probabilidade de ignición e a súa velocidade de propagación. Debe terse en conta esta situación á hora de realizar a extinción dun incendio, pois en caso contrario o incendio podería extenderse de forma que a brigada quede atrapada. Por iso é de capital importancia establecer e manter en todo momento unha ruta de escape pola que abandoar a zona en caso de ser necesario. Os destinos de ditas rutas deben ser lugares nos que non exista perigo de ser alcanzados polo lume. Tamén se debe ter en conta que, segundo evolucione o incendio, as rutas poden cambiar, polo que se deben re-evaliar de forma periódica.

Co obxetivo de apoiar esta tarefa desenvolveuse, no marco do proxecto Civil UAVs Initiative (CUI) da Xunta de Galicia e a empresa Avincis Aviation Spain SA, un algoritmo que calcula rutas de escape, a pe, para a evacuación dos medios de extinción terrestres.

2. ALGORITMO DE RUTAS DE ESCAPE

O algoritmo, desenvolto na súa totalidade en R, centrase no cálculo das rutas de escape mais rápidas para brigadas a través do terreno, baseándose na pendente, o modelo de combustible e as redes viarias presentes no área do incendio. Como datos de entrada obrigatorios o usuario debe proporcionar a xeometría do perímetro do incendio, a posición de partida da brigada, a capa de modelos de combustible, a dirección de avance do incendio e os posibles puntos de destino. De forma resumida, a metodoloxía proposta para o cálculo da ruta segue os seguintes pasos:

1. Obtense o Modelo Dixital do Terreo (MDT), as vías de transporte, os cursos de auga e as augas estancadas na rexión do incendio. Estos datos obteñense dos servizos da Infrastructura de Datos Espaciais de España (IDEE).

2. Procesa os datos obtidos no paso anterior. En concreto, calcula a pendente do terreo, utilizando o MDT, e rasteriza as vías de transporte, os cursos de auga e as augas estancadas.

3. Cárganse os datos do modelo de combustible na zona. Estos datos foron proporcionados por Avincis e clasifican o terreo según se o tipo de combustible son pastos, matorral baixo, matorral alto, bosques con continuidade vertical ou bosque con discontinuidade vertical. En caso de dispoñer información sobre a clase de tramos de carreteras, establece que nos tramos de autopistas, carreteras, autovías, carriles bici e urbanos non se encontra ningunha clase de combustible. Esta información pode introducirse mediante un arquivo `csv` en caso de utilizarse os datos de carreteras do IDEE, xa que non contan con esta información.

4. Filtranse os datos á zona contraria ao avance do incendio. Isto permite alixerar os cálculos posteriores e garantiza a seguridade da ruta final.

5. Xérase un grafo dirixido que relaciona cada píxel cos píxeles adxacentes, utilizando a libraría *igraph*. A cada arista do grafo lle corresponde como peso o tempo que as brigadas tardan en transitar entre cada un dos nodos. Este tempo depende do tipo de combustible polo que se transita e da pendente do terreo. Considerase que as masas de auga e o perímetro do incendio son intransitables, salvo nos lugares polos que transcurra algún camiño.

6. Finalmente, cálculase, , para cada posible destino definido polo usuario, a ruta mais rápida utilizando o algoritmo de Dijkstra[1] ([2]).

Para o cálculo do tempo de transito en cada arista (paso 5 da metodoloxía) considerase que a velocidade dunha brigada, co seu equipamento é de 6,4 km/h, baixo condicións ideais, é dicir, cunha pendente do 0 % e sen presenza de ningunha clase de combustible. Esta velocidade vese afectada por calquera alteración nos parámetros de pendente e combustible, afectando de forma porcentual á velocidade de desprazamento, considerando cada un dos factores de forma independente. Por exemplo, en caso de ter unha pendente ascendente de entre un 40 % e un 70 %, a velocidade disminuirá un 50 % polo que pasará a 3,2 km/h. Se además hai presenza de combustible tipo pasto, a velocidade baixará outro 25 %, polo que a velocidade nesa zona será de 2,4 km/h. Estos factores de corrección se encontran tabulados e foron proporcionados por Avincis.

Ademais dos datos de entrada obrigatorios, o algoritmo admite outros parámetros opcionais. Por exemplo, permite utilizar datos almacenados en local en lugar de utilizar os dispoñibles a través de servizos, establecer unha distancia de seguridade respecto ao perímetro do incendio e unha cota cota superior á pendente que se pode transitar. O usuario tamén pode definir de obstáculos de forma manual.

O algoritmo fai uso dalgunhas das librarías mais utilizadas para estatística espacial, como son *sf* ([5]), *terra* ([4]) e *tidyterra* ([3]). Tamén fai uso doutras librarías como a libraría *igraph* ([1]), para xerar o grafo e aplicar o algoritmo de Dijkstra e a libraría *dplyr* ([6]), que facilita o procesado dos datos (especialmente no cálculo do tempo de cada arista do grafo). Para facilitar o seu uso e despliegue, algoritmo utilizase nun entorno creado con docker, e introducindo os datos de entrada mediante un arquivo `JSON`. Docker é unha plataforma de software que permite empaquetar software en contenedores, que inclúen todo o necesario para que este se poida executar. De esta forma, créase un entorno illado, que non se ve influenciado pola máquina na que se esté executando. Esto permite xerar un entorno estable que permita evitar problemas de compatibilidade que poden darse pola instalación de outras versións das librarías utilizadas.

Na Figura 1 pode verse un exemplo de execución para un incendio en Galicia. Neste caso, estableceuse que a ruta non debe transcorrer a menos de 100 metros do perímetro do incendio. Como o

[1] Determina, dado un nodo de orixe e outro de destino nun grafo, o camiño curto ou con mínimo custo, en caso de ser un grafo ponderado con pesos positivos (esto pode ser, por exemplo, o tempo de transito entre nodos).

punto de partida se sitúa a menos de dita distancia, a ruta trata de afastarse o mais rápido posible do perímetro do incendio, posto que aínda que non está estrictamente prohibido desprazarse nesa zona, sí está fortemente penalizado o paso por ela. Cara o final, a ruta transcorre por unha carretera, posto que desa forma evitase a penalización de velocidade por presenza de combustible.

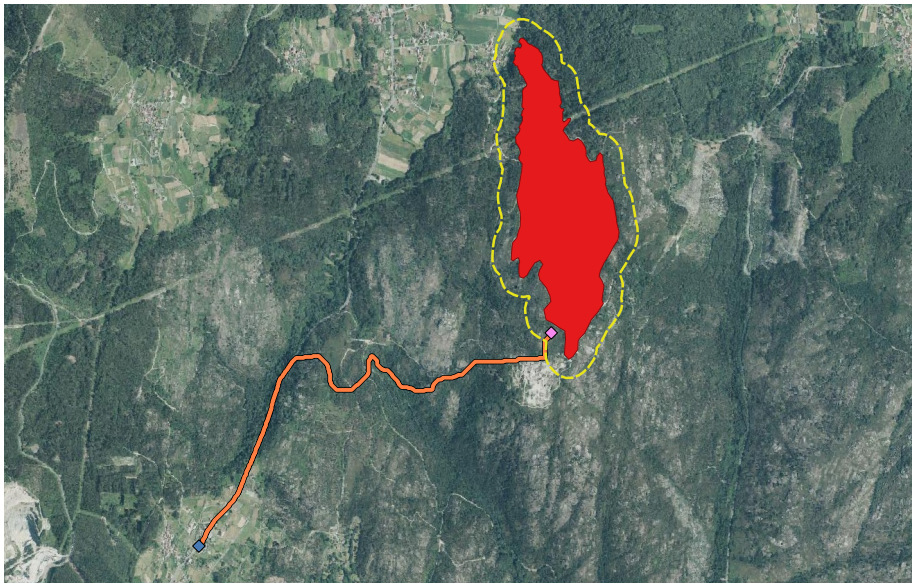


Figura 1: Ruta de escape estimada polo algoritmo (naranja), o perímetro do incendio (vermello), o punto de partida (rosa), o punto de destino (azul) e un buffer de 100 metros do perímetro do incendio (amarelo).

3. CONCLUSIÓNS

A solución aportada permite ao usuario deste algoritmo obter unha ruta de escape dende a posición actual da brigada a unha zona segura. Utilizando os datos de entrada o algoritmo é capaz de proporcionar unha ruta de escape tendo en conta multitude de carácterísticas do terreo como poden ser a pendente do terreo, o modelo de combustible e a existencia de vías, ríos e aguas estancadas. Ademais, o usuario pode definir obstáculos presentes no terreo que a brigada non sería capaz sortear ou establecer outras condicións como a distancia de seguridade respecto ao incendio.

Ao executarse nun entorno docker, o algoritmo pode executarse dende outros equipos sen perigo de sufrir problemas por incompatibilidades do sistema. Dende o punto de vista computacional, o algoritmo non ten grandes problemas, pois nas probas de rendemento o algoritmo foi capaz de calcular varias rutas en menos dun minuto e medio, considerando unha rexión definida por un buffer de 3 kilómetros do perímetro do incendio. Neste aspecto, o unha posible mellora do algoritmo sería poder traballar con datos en tempo real, posto que agora as posicións e a dirección de avance se deben proporcionar de forma manual.

AGRADECEMENTOS

Os investigadores Marta Rodríguez, Manuel Antonio Novo, e María José Ginzo agradecen o apoio do proxecto CUI da Axencia Galega de Innovación (GAIN) da Xunta de Galicia e á empresa Avincis Aviation Spain SA.

Referencias

[1] Csárdi G, Nepusz T, Traag V, Horvát Sz, Zanini F, Noom D, Müller K (2024). *igraph: Network Analysis and Visualization in R*. R package version 2.0.3. https://CRAN.R-project.org/package=igraph

[2] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1, 269-271. doi: https://doi.org/10.1007/BF01386390

[3] Hernangómez D. (2023). Using the tidyverse with terra objects: the tidyterra package. Journal of Open Source Software, 8(91), 5751. https://doi.org/10.21105/joss.05751

[4] Hijmans R. (2024). terra: Spatial Data Analysis. R package version 1.7-71. https://cran.r-project.org/web/packages/terra/index.html

[5] Pebesma E. (2024). Sf: Simple Features for R. R package versión 1.0-16. https://cran.r-project.org/web/packages/sf/index.html

[6] Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.2. https://CRAN.R-project.org/package=dplyr

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

El análisis por componentes en las ciencias sociales

Jorge Alejandro Obando[1], Aura Viviana Rincón Ramirez[2] Laura Nathalia Obando[3]
[1] universidad Cooperativa de Colombia
[2] universidad Cooperativa de Colombia
[3] universidad Cooperativa de Colombia

RESUMO

El Análisis por Componentes Principales (ACP) en las ciencias sociales es una herramienta valiosa para estudiar el impacto del conflicto en el tejido social. En el contexto de El Castillo Meta, una zona afectada por el conflicto armado, se utilizó una encuesta de cohesión social aplicada a 405 personas en eventos comunitarios. El ACP permitió identificar cómo actividades como talleres, eventos y promoción están interrelacionadas y son esenciales para la reconstrucción social. Los resultados mostraron que variables como la participación y la reconciliación tienen una fuerte correlación, indicando que estas actividades son fundamentales para la memoria histórica y la cohesión social. Concluye que el ACP es crucial para visualizar la resiliencia y transformación social en contextos de posconflicto, facilitando una comprensión profunda de las dinámicas sociales y comunitarias.

Palabras e frases chave: R en ciencias sociales, tejido social, ACP

1. INTRODUCIÓN

El concepto de tejido social en contextos de posconflicto se refiere a la reconstrucción de las relaciones y estructuras comunitarias que han sido fragmentadas por la violencia. Este proceso implica el fortalecimiento de las redes de apoyo mutuo y la cohesión social [1]. Las prácticas pedagógicas en zonas de conflicto también contribuyen significativamente a este proceso, al fomentar la resiliencia y la capacidad de los individuos para enfrentar y superar las adversidades [2]. Los foros de escucha y otros acercamientos a la comunidad para la pacificación y la reconciliación intentan incorporar las voces de los afectados por la violencia para construir una agenda pública de seguridad y paz [3]. Este esfuerzo es un reflejo de iniciativas para la reconstrucción del tejido social a través de la participación comunitaria y el diálogo, por ejemplo, las brigadas y grupos sociales contribuyen a la cohesión social y la resistencia [4]. El tejido social en El Castillo, Meta, fue profundamente afectado por la violencia del conflicto armado, especialmente durante el periodo de 2002 a 2008. Este proceso implicó el desplazamiento forzado y el vaciamiento de comunidades, resultando en la desintegración de las estructuras sociales y económicas locales [5].

Para examinar este proceso de reconstrucción del tejido social en El Castillo Meta, se aplica una encuesta relacionada con la cohesión social a 405 personas ubicadas en eventos sociales, como talleres, reuniones, peregrinaciones y fechas especiales donde la comunidad se reúne para hacer memoria del conflicto y tomar decisiones que propician resiliencia apoyando la reconstrucción del tejido social. la encuesta es una escala Likert con 14 variables a la cual se le aplica el análisis por componentes.

El análisis por componentes principales (PCA), ha demostrado ser una herramienta estadística importante en diversas áreas de las ciencias sociales, por ejemplo, En un estudio reciente, [6] exploró el uso del PCA con correlaciones policóricas para evaluar el consumo de medios y la participación política en Argentina, Chile y Uruguay, demostrando que los coeficientes policóricos ofrecen mayor precisión en variables ordinales. Así mismo se encontró que para que los estudiantes de posgrado en ciencias sociales dominen métodos cuantitativos avanzados para competir en el mercado laboral ell PCA tiene lacapacidad para proporcionar insights significativos en diferentes contextos [7].

2. Desarrollo del Análisis por componentes (PCA)

La tabla muestra los valores del MSA (Measure of Sampling Adequacy) tanto de manera general como para cada una de las variables individuales relacionadas con la participación en actividades de memoria histórica en El Castillo, Meta. El MSA es una medida que indica la adecuación de las variables para ser incluidas en un análisis factorial o de componentes principales, con valores cercanos a 1 indicando una mayor adecuación.

	MSA
MSA General	0.843
Actividades	0.867
Eventos	0.849
Talleres	0.869
Reconciliación	0.888
Promoción	0.912
Iniciativas	0.852
Construcción	0.807
Participación	0.829
Motivación	0.859
Interés	0.843
Contribución	0.814
Transformación	0.846
Identificación	0.760
Recorridos	0.731

Tabla 1. Contraste de Kaiser-Meyer-Olkin

El valor general del MSA es 0.843, lo que sugiere que, en promedio, las variables son adecuadas para el análisis. Individualmente, todas las variables presentan valores de MSA por encima de 0.7, con "Promoción" (0.912) y "Reconciliación" (0.888) siendo las variables con mayor adecuación, lo que indica que son particularmente apropiadas para este tipo de análisis. Por otro lado, "Recorridos" (0.731) y "Identificación" (0.760) tienen los valores más bajos, aunque aún se consideran adecuadas.

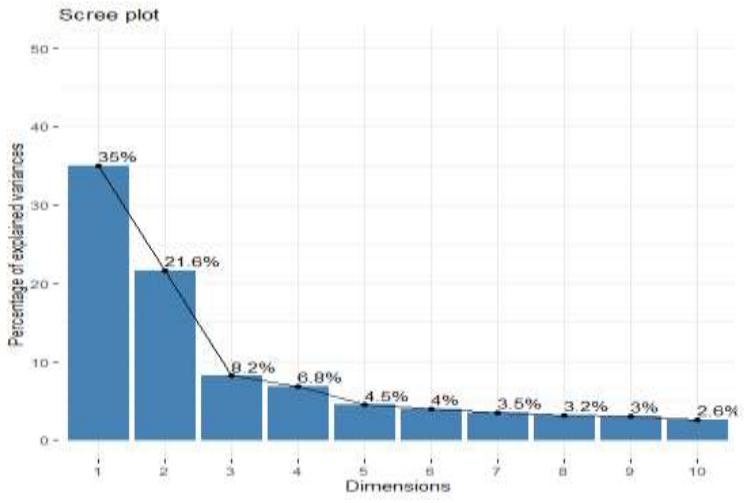


Figura 1. scree plot.

Esta observación dada en la figura 1, subraya la eficiencia del PCA para identificar los factores más significativos y resalta la importancia de determinar el número adecuado de componentes para un análisis preciso y útil en estudios de ciencias sociales. Por ejemplo, del segundo al tercer componente se nota una disminución considerable de varianza, presentándose el codo, momento óptimo para detener el análisis, ya que los

componentes adicionales aportan una explicación marginalmente menor de la varianza total, entonces para este análisis se tomas dos dimensiones que aportan el 56,6% de la varianza total.

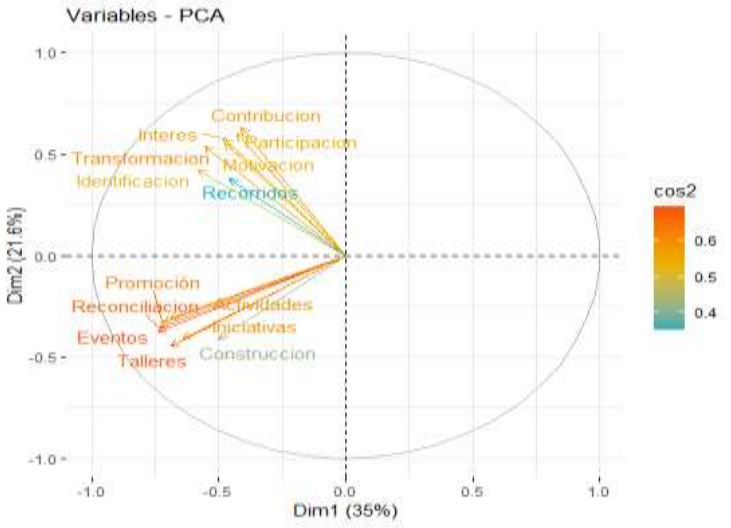


Figura 2. Biplot de análisis de componentes principales

La figura muestra un análisis de componentes principales (PCA) de las variables relacionadas con la participación en actividades de memoria histórica en el contexto del tejido social en El Castillo, Meta. El PCA, que explica un 35% y un 21.6% de la varianza total en los dos primeros componentes, permite visualizar cómo las actividades como eventos, talleres, y promoción están fuertemente correlacionadas y agrupadas en el gráfico. Esto sugiere que estas actividades son esenciales para la reconstrucción del tejido social y la memoria histórica en la comunidad. Además, variables como participación, motivación, interés y transformación indican una fuerte relación entre la percepción individual de la importancia de estas actividades y su impacto en la transformación social.

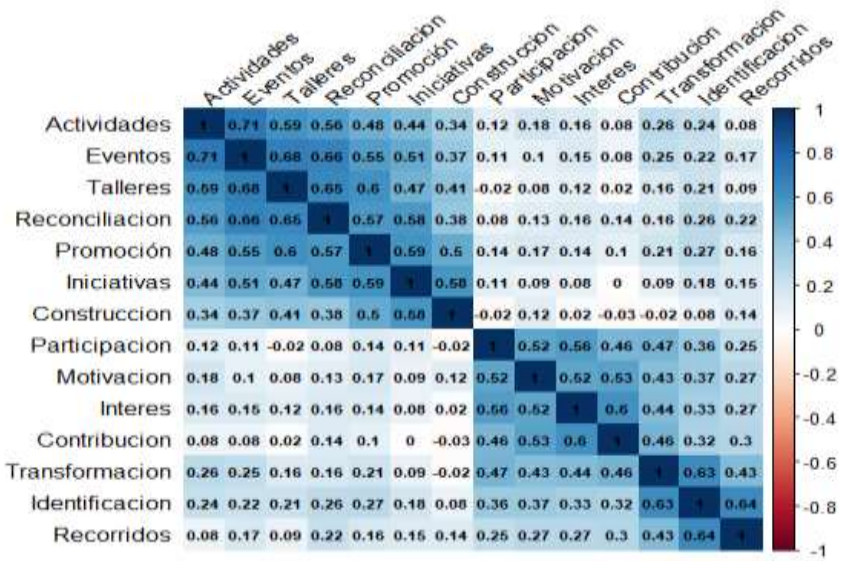


Figura 3. Mapa de calor

La figura 2 un mapa de calor de la matriz de correlaciones entre diversas variables relacionadas con la participación en actividades de memoria histórica en el contexto del tejido social en El Castillo, Meta. Se observa que variables como "Actividades", "Eventos", "Talleres" y "Reconciliación" tienen fuertes correlaciones positivas entre sí, lo que sugiere que la participación en una de estas actividades está altamente asociada

con la participación en las otras. Esto refuerza la importancia de las actividades colectivas en la reconstrucción del tejido social.

3. CONCLUSIÓNS

El uso de R en las ciencias sociales permite un análisis profundo y detallado de variables críticas como tejido_social, cohesión, posconflicto y resiliencia. Mediante técnicas avanzadas como el Análisis por Componentes Principales (ACP), R facilita la identificación de patrones y relaciones clave entre estas variables, proporcionando insights valiosos para la reconstrucción del tejido social en contextos de posconflicto. La capacidad de R para manejar grandes conjuntos de datos y realizar análisis complejos lo convierte en una herramienta indispensable para investigadores y académicos en el campo de las ciencias sociales.

Referencias

[1] Cataño, S. V., Jiménez, E. A., & López, M. (2023). La trayectoria de quienes quedan. Narrativa y desaparición forzada en Colombia. *Textos y Contextos*, 27, e4322.

[2] HUMAN Review. (2023). Prácticas pedagógicas para la reconstrucción del tejido social en territorios de conflicto. *HUMAN Review*.

[3] Hernández González, G. (2020). Uso de la categoría tejido social/comunitaria en los Foros Escucha para la Pacificación y la Reconciliación Nacional. *Revista Kavilando*, 12(2), 429-439.

[4] Páez, C. (2013). La práctica de la resistencia en las brigadas muralistas de los '80. *Universidad de Chile*.

[5] Centro Nacional de Memoria Histórica. (2015). *Pueblos arrasados: Memorias del desplazamiento forzado en El Castillo (Meta)*. Bogotá: CNMH.

[6] González-Bustamante, B. (2023). Análisis de Componentes Principales con correlaciones policóricas: Aplicación en consumo de medios.

[7] Vijverberg, W. P. (1997). The quantitative methods component in social sciences curricula in view of journal content. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 16(4), 621-62

Anexo

objetos de R	Código R Usado para PCA
Librerias	library(factoextra), library(ggplot2), library(readxl),library(corrplot)
Lectura del archivo Excel desde la ruta especificada	data <- read_excel("D:/Academicos2024/EventoGalicia/Encuesta.xlsx")
Estandarizar los datos	data_scaled <- scale(data)
PCA	pca_result <- prcomp(data_scaled, center = TRUE, scale. = TRUE)
Scree plot	fviz_screeplot(pca_result, addlabels = TRUE, ylim = c(0, 50))
Círculo de correlación	fviz_pca_var(pca_result, col.var = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
Matriz de correlaciones	cor_matrix <- cor(data)
Gráfico de calor con la matriz de correlaciones	corrplot(cor_matrix, method = "color", tl.col = "black", tl.srt = 45, addCoef.col = "black", number.cex = 0.7)

Tabla 2. Códigos en R para la realización de un PCA

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

Efecto da asma na calidade de vida autopercibida polos pacientes en España

Alba Paz-Castro[1,2], José Manuel Amoedo[3]

[1] Grupo Bioloxía do Linfocito (BioLinfo) Departamento de Bioquímica e Bioloxía Molecular, Facultade de Bioloxía- Centro de Investigacións Biolóxicas (CIBUS), Universidade de Santiago de Compostela, Santiago de Compostela, España.

[2] Grupo de Investigación Traslacional en Enfermidades das Vías Respiratorias (TRIAD), Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, España.

[3] Grupo de Investigación ICEDE, Departamento de Economía Aplicada, Facultade de Ciencias Económicas e Empresariais, Universidade de Santiago de Compostela

RESUMO

O asma é unha enfermidade respiratoria crónica que afecta significativamente á calidade de vida dos doentes. Este traballo ten como obxectivo analizar o impacto da asma na calidade de vida autopercibida dos pacientes en España, empregando datos provintes das enquisas de saúde nacionais. A través de Propensity Scoore Matching e de estimacións econométricas loxísticas estimouse a influencia do asma na calidade de vida dos pacientes asmáticos.

Empregáronse os datos dispoñibles na enquisa estandarizada de saúde do Instituto Nacional de Estadística (INE) do ano 2020 onde se mediron a percepción dos pacientes sobre o seu estado de saúde e benestar xeral. As dimensións avaliadas inclúen o impacto físico, emocional e social da asma, permitindo unha visión global do problema.

Para levar a cabo a análise emprégase tres librerías de R que permiten levar a cabo as diferentes fases da análise. En primeiro lugar, lévase a cabo unha Análise de Compoñentes Principais (ACP) que permite simplificar as covariables de control para evitar posibles problemas de multicolinealidade. En segundo lugar, a librería *MatchIt,* permite obter grupos de tratamento (doentes con asma) e de control (doentes sen asma) similares mediante o uso de Propensity Score Matching. Finalmente, a librería *nnet* permite realizar estimacións loxísticas multinomiais para analizar como afecta a asma ós doentes que a padecen.

Os resultados preliminares indican unha correlación significativa entre a presencia de asma no paciente e unha menor calidade de vida. Neste senso, as estimacións loxístixas multinomiais indican que o feito de padecer asma incrementa de forma significativa a probabilidade de empeorar a saúde autopercibida. Ademais, os resultados amosan como a probabilidade de que a saúde autopercibida empeore ó padecer asma é especialmente elevada entre os individuos con peor saúde.

Palabras e frases chave: asma, calidade de vida, saúde, benestar, España

XI Xornada de Usuarios de R en Galicia
Santiago de Compostela, 24 de outubro do 2024

CALCULANDO A SEVERIDADE DUN INCENDIO FORESTAL CON R

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

[1]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga)
[2]Departamento de Estatística, Análise Matemática e Investigación Operativa, Universidade de Santiago de Compostela

RESUMO

O algoritmo desenvolto permite calcular a severidade dun incendio unha vez extinto, proporcionando información sobre as posibilidades de rexeneración da vexetación afectada. Para isto, utilízase un índice denominado Normalized Burn Ratio (NBR) que se calcula a partir de imaxes obtidas do satélite Sentinel-2, mediante o servizo proporcionado por Google Earth Engine. Empregando librerías de R de estatística espacial como *terra*, *sf* ou *tidyterra*, o algoritmo desenvolto conecta cun script de Python capaz de descargar as imaxes de satélite previas e posteriores ó incendio, e a continuación procesa estas imaxes para calcular a diferencia dos índices Normalized Burn Ratio e clasificar os valores de cada píxel indicando a severidade do incendio nos mesmos e a probabilidade de rexeneración. Isto permite coñecer as zonas nas que é necesario intervir para axudar á rexeneración da vexetación.

Palabras e frases chave: Incendios forestais, rexeneración, dNBR, GEE, imaxes satélite.

1. INTRODUCIÓN

Tras un incendio forestal, a capacidade de rexeneración da vexetación depende de diversos factores como a severidade do incendio, as especies vexetais presentes, as condicións ambientais... (Sterner *et al.*, 2022). Un dos obxectivos despois dun incendio forestal debe ser recuperar o ecosistema orixinal previo á existencia do incendio. A vexetación nativa da zona a miúdo adáptase para rexenerarse ou reproducirse despois dun incendio forestal de baixa ou media intensidade, pero en incendios de alta severidade as áreas queimadas tardan en rexenerar.
Unha vez que se extingue un incendio forestal todos os esforzos deben centrarse na rehabilitación da zona e a estabilización do solo. O primeiro paso para isto é a avaliación da severidade do incendio, que se define como o grado no que unha zona foi alterada polo lume. A creación dun mapa que reflicta os efectos do incendio na superficie do solo e o estado deste, facilita a identificación das áreas potenciais de preocupación e o recoñecemento inicial do terreo. Isto axuda na toma de decisión das áreas nas que os tratamentos de rehabilitación poden ser máis eficaces.
O obxectivo do algoritmo desenvolvido é proporcionar apoio para esta tarefa, calculando un mapa que represente a severidade dun incendio. Para isto, utilizaranse imaxes de satélite para calcular o Normalized Burn Ratio (NBR) do terreo previo ó incendio e despois deste. O NBR é un índice que se utiliza comunmente para identificar zonas queimadas calculando a reflectancia das bandas de infravermellos. Un valor alto do NBR indica vexetación saudable, mentres que un valor baixo indica solo descuberto e áreas recentemente queimadas. A diferenza entre os valores do índice, Differenced Normalized Burn Ratio (dNBR) é a diferencia entre o NBR antes e despois do incendio e proporciona información de gran utilidade sobre a recuperación da vexetación. Os valores positivos do dNBR indican zonas que sufriron cambios debido a un incendio, e os valores negativos indican áreas de rexeneración ou crecemento da vexetación despois do incendio. Os valores do dNBR

poden variar polo que a súa interpretación debe levarse a cabo mediante unha avaliación de campo, sempre que sexa posible. Porén, o Servizo Xeolóxico dos Estados Unidos (USGS) desenvolveu unha clasificación dos valores do dNBR para interpretar a gravidade dun incendio (Key & Benson, 2006), tal e como se amosa na Figura 1.

Severity Level	dNBR Range (scaled by 10^3)	dNBR Range (not scaled)
Enhanced Regrowth, high (post-fire)	-500 to -251	-0.500 to -0.251
Enhanced Regrowth, low (post-fire)	-250 to -101	-0.250 to -0.101
Unburned	-100 to +99	-0.100 to +0.99
Low Severity	+100 to +269	+0.100 to +0.269
Moderate-low Severity	+270 to +439	+0.270 to +0.439
Miderate-high Severity	+440 to +659	+0.440 to +0.659
High Severity	+660 to +1300	+0.660 to +1.300

Figura 1: Clasificación da severidade dun incendio a partir dos valores do dNBR, creada polo USGS.

2. CÁLCULO DA SEVERIDADE DO INCENDIO

O algoritmo desenvolto con R proporciona un mapa co dNBR clasificado do incendio, comparando o NBR actual da zona cos valores previos ó incendio. Utiliza a clasificación do USGS (Figura 1), resaltando as zonas sobre as que hai que actuar xa que non se producirá unha rexeneración natural. En primeiro lugar, o algoritmo obtén da base de datos o perímetro final asociado ó incendio a partir do identificador introducido polo usuario. Este perímetro é necesario para establecer a rexión de interese na que se calcula o dNBR. O seguinte paso é a descarga de imaxes previas e posteriores ó incendio, para o que se utilizan as imaxes do Sentinel-2 mediante o servizo proporcionado por Google Earth Engine (GEE). Para conectar con este servizo emprégase un script de Python que se executa dende R.

Unha vez que se dispón das imaxes, hai que filtrar aquelas que son de baixa calidade. A continuación, agréganse todas as imaxes previas ó incendio, creando unha imaxe que é a composición de todas usando os valores medianos. Faise o mesmo coas imaxes posteriores. Deste xeito, tense unha única imaxe previa ó incendio e outra posterior. Calcúlase o NBR de cada unha das imaxes, e despois réstase para obter o valor do dNBR previamente mencionado.

Na imaxe resultante, clasifícanse os valores do dNBR en función da clasificación proporcionada polo USGS que se amosa na Figura 1, e proporciónase este mapa clasificado como saída. Deste xeito, o usuario pode ver rapidamente as zonas nas que é necesaria a intervención humana.

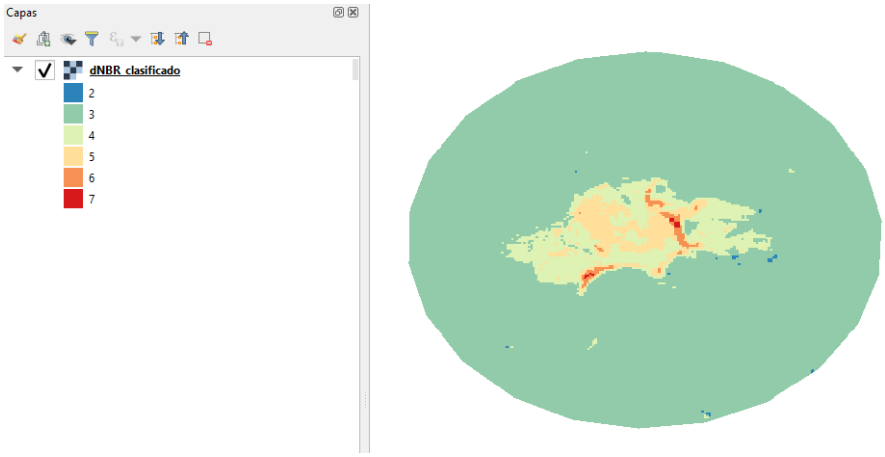


Figura 2: Severidade dun incendio segundo a clasificación do USGS. As cores fan referencia a esta clasificación, no que os valores máis altos indican maior severidade do incendio.

Na Figura 2 amósase un exemplo da saída do algoritmo vista desde o visor de QGIS. Trátase do dNBR clasificado na zona na que se produciu un incendio forestal, o que permite ver de xeito rápido

as zonas nas que se debe actuar para favorecer a rexeneración. Neste caso, as zonas de cor laranxa son de risco alto de non rexeneración, mentres que as zonas de cor vermella indican que se debe actuar de xeito inmediato, xa que non se vai producir unha rexeneración natural da vexetación.
Para o tratamento das imaxes de satélite e a creación dos mapas, o algoritmo emprega librarías para datos espaciais de R como son *terra*, *sf* ou *tidyterra*.

3. CONCLUSIÓNS

A solución aportada por este algoritmo permite establecer unha estimación dos danos provocados por un incendio forestal en función dos valores do dNBR e seguindo unha clasificación establecida polo USGS.
É unha ferramenta que pode ser de gran axuda en canto ó restablecemento do solo e a vexetación, xa que proporciona dun xeito rápido as zonas nas que se debería intervir con traballos de rexeneración xa que esta non se vai producir de forma natural.
O algoritmo pode ser executado en diferentes momentos despois dun incendio, o que permite analizar a evolución da rexeneración do solo, iniciando as labores de rexeneración tan pronto como sexan necesarias. Pode ocorrer que inmediatamente despois dun incendio unha zona non mostre signos de que sexa necesaria a actuación humana, pero co paso do tempo, e analizando a rexeneración da vexetación segundo os valores do dNBR, se detecte que a evolución da vexetación non sexa a esperada e se decida intervir.
Esta ferramenta non substitúe o traballo de campo, no que se estuda e analiza o solo despois dun incendio. Porén, proporciona unha axuda a estes traballos xa que, en zonas nas que hai moitos incendios, pode servir como primeiro indicador para avaliar as zonas nas que se deben centrar os estudos e os esforzos de rexeneración.
Este algoritmo foi desenvolto en colaboración coa empresa de aviación Avincis, no marco do proxecto *Civil UAVs Initiative* da Axencia Galega de Innovación (Xunta de Galicia).

Referencias

[1] Hernangómez, D. (2023). Using the tidyverse with terra objects: the tidyterra package. *Journal of Open Source Software*, 8(91), 5751. ISSN 2475-9066.

[2] Key, C.H., Benson, N.C. (2006). Landscape Assessment (LA). In Lutes, D.C., Keane, R.E., Caratti, J.F., Key, C.H., Benson, N.C., Sutherland, S., Gangi, L.J. (eds.): *FIREMON: Fire effects monitoring and inventory system.* USDA Forest Service, Rocky Mountain Research Station. Gen. Tech. Rep. RMRSGTR-164-CD, 1-5

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

[4] Pebesma, E., Bivand, R. (2023). *Spatial Data Science*. Chapman and Hall, New York

[5] Sterner, S., Aslan, C., Chaudhry, T. (2022). Forest management effects on vegetation regeneration after a high severity wildfire: A case study in the southern Cascade range. *Forest Ecology and Management*, 520, 120394.

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

Análisis de nubes de puntos masivas en R

Nataly Romarís-Lodeiro[1], Olamar Benavente-Fernández[1], Rubén Fernández-Casal[2] y Salvador Naya[2]

[1]Centro Mixto de Investigación UDC-Navantia, Universidade da Coruña.
[2]Grupo MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña.

RESUMO

Hoy en día están disponibles distintas técnicas de teledetección que permiten obtener mediciones de alta resolución de la superficie de estructuras tridimensionales. Entre ellas podríamos destacar las mediciones LiDAR (Light Detection and Ranging) obtenidas mediante pulsos láser. Las nubes de puntos resultantes tienen aplicación en muchos campos, como por ejemplo ciencias forestales, morfología o control de procesos de fabricación. En este trabajo se pretende hacer una breve revisión de algunas de las herramientas disponibles en R para la manipulación y el análisis de este tipo de datos.

Palabras e frases chave: datos LiDAR, isosuperficies, mallas triangulares, coordenadas espaciales 3D.

1. INTRODUCIÓN

Cada observación de una nube de puntos LiDAR incluye coordenadas espaciales en tres dimensiones y opcionalmente otros atributos, como la intensidad del pulso reflejado. El primer paso para el análisis de este tipo de datos es su importación en R y su representación gráfica. Para ello, la recomendación es emplear el paquete *rgl* (Murdoch y Adler, 2023), especialmente si se trata de conjuntos de datos de gran tamaño. Este paquete incluye funciones para importar, construir y visualizar objetos tridimensionales (incluyendo nubes de puntos y mallas triangulares), además de muchas otras operaciones geométricas (como realizar transformaciones afines o segmentar un objeto en partes orientadas en una dirección determinada).

También pueden resultar de utilidad paquetes de R diseñados para su aplicación en campos específicos. Entre ellos podemos destacar los paquetes *lidR* (Roussel y Auty, 2023) y *lasR* (Roussel, 2024) para el análisis de datos forestales, y el paquete *Morpho* (Schlager, 2017) para el análisis de datos morfológicos mediante puntos de referencia y mallas de superficie.

El conjunto de datos puede estar formado por decenas o cientos de millones de puntos por lo que antes de realizar cualquier análisis resulta necesario reducir su dimensión. Para ello se pueden emplear distintas técnicas de discretización, entre ellas binning o voxelización, utilizando los paquetes *npsp* (Fernández-Casal, 2024) o *VoxR*

(Lecigne et al., 2018), por ejemplo. Aun así, los requerimientos en cuanto a la precisión de los resultados, pueden hacer necesario el trabajar con nubes formadas por millones de puntos.

En muchas ocasiones las nubes de puntos no son una representación adecuada del objeto escaneado e interesa aproximar su forma mediante una malla triangular. Para ello se pueden emplear diversas técnicas, entre ellas la envoltura α-convexa de la nube de puntos, niveles de densidad (isosuperficies) o el método Ball-Pivoting.

Para calcular la envoltura α-convexa de un conjunto de puntos se puede emplear el paquete *alphashape3d* (Lafarge et al., 2014), o el paquete *alphahull* (Pateiro-López y Rodríguez-Casal, 2010) para conjuntos de puntos en el plano. Sin embargo, como estos métodos están basados en la triangulación de Delaunay, sus requerimientos computacionales son muy elevados y en la práctica solo pueden ser empleados con unos pocos miles de datos.

Otra técnica para aproximar la forma del objeto, especialmente útil si hay errores de medición, es la generación de isosuperficies correspondientes a niveles de densidad (ver e.g. Chacón y Duong, 2018, Sección 6.1). Este enfoque se basa en estimaciones de la densidad multidimensional en una rejilla regularmente espaciada, que pueden ser obtenidas mediante suavizado tipo núcleo, empleando los paquetes *ks* (Duong, 2022) o *npsp*, entre otros. También se podría emplear directamente la rejilla binning para acelerar los cálculos. La isosuperficie estará determinada por un determinado valor (nivel) de densidad y serviría como una aproximación de la forma del objeto escaneado. Para obtener la correspondiente malla triangular se puede emplear el paquete *misc3d* (Feng y Tierney, 2008). Este paquete genera isosuperficies tridimensionales utilizando una versión optimizada del algoritmo Marching Cubes (Lorensen y Cline, 1987). El objeto resultante puede representarse utilizando el paquete *rgl* o los gráficos estándar de R.

También se puede emplear el algoritmo Ball-Pivoting, implementado en la librería *Rvcg* (Schlager, 2017), para la generación de una malla de triángulos a partir de la nube tridimensional de puntos. El mecanismo de este algoritmo es relativamente sencillo; se considerará que tres puntos forman un triángulo de la malla siempre y cuando una bola de radio p (especificado previamente por el usuario) los toca sin incluir ningún otro punto de la nube. Este algoritmo partirá de un triángulo inicial o triángulo "semilla" girando la bola alrededor de los bordes hasta encontrar nuevos puntos para formar más triángulos repitiendo este proceso hasta que se hayan probado todos los bordes accesibles y todos los puntos hayan sido considerados. En nubes de puntos con densidades desiguales, es posible volver a ejecutar el algoritmo con un radio mayor para evitar que quede una malla no hermética.

Además, el paquete *Rvcg* implementa métodos avanzados para manipular mallas de superficies y realizar análisis de formas. Entre ellas, permite la limpieza de la malla, por ejemplo, eliminando piezas aisladas, seleccionadas por una cantidad mínima de caras o de un diámetro por debajo de un umbral determinado. También permite realizar suavizados, calcular vectores normales, calcular distancias geodésicas y el submuestreo de la superficie de la malla triangular, devolviendo un conjunto de puntos ubicados en esa malla, entre otras funciones.

Por último, es posible desarrollar procedimientos adicionales empleando herramientas implementadas en Python, y que pueden ser llamadas desde R empleando el paquete *reticulate* (Ushey et al., 2024). Esto también permite mejorar computacionalmente ciertas operaciones que en R todavía presentan limitaciones.

AGRADECEMENTOS

Este trabajo ha sido financiado por la Axencia Galega de Innovación (GAIN) de la Xunta de Galicia y la empresa Navantia (SEPI), en el marco del Centro Mixto de Investigación UDC-NAVANTIA, con el proyecto "O estaleiro do futuro" (IN853C).

Referencias

Chacón J.E., Duong T. (2018). *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC.

Bernardini F., Mittleman J., Rushmeier H., Silva C., Taubin G. (1999). The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics*, 5(4), 349-359.

Duong T. (2022). *ks: Kernel Smoothing*. R package version 1.14.0, https://CRAN.R-project.org/package=ks.

Feng D., Tierney L. (2008). Computing and Displaying Isosurfaces in R. *Journal of Statistical Software*, 28, 1-24.

Fernández-Casal R. (2024). *npsp: Nonparametric Spatial Statistics.* R package version 0.7-14. https://rubenfcasal.github.io/npsp.

Lafarge T., Pateiro-López B., Possolo A., Dunkers J. (2014). R Implementation of a Polyhedral Approximation to a 3D Set of Points Using the α-Shape. *Journal of Statistical Software*, 56, 1-19.

Lecigne B., Delagrange S., Messier C. (2018). Exploring trees in three dimensions: VoxR, a novel voxel-based R package dedicated to analysing the complex arrangement of tree crowns. *Annals of Botany*, 121, 589-601.

Lorensen W.E., Cline H.E. (1987). Marching Cubes: A High Resolution 3D Surface Reconstruction Algorithm. *Computer Graphics*, 21(4), 163-169.

Murdoch D., Adler D. (2023). *rgl: 3D Visualization Using OpenGL*. R package version 1.1.3, https://CRAN.R-project.org/package=rgl.

Pateiro-López B., Rodríguez-Casal A. (2010). Generalizing the convex hull of a sample: the R package alphahull. *Journal of Statistical Software*, 34, 1-28.

Roussel J.R., Auty D. (2023). *lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. R package version 3.1.0. https://cran.r-project.org/package=lidR.

Roussel, J.R. (2024). lasR: Fast and Pipeable Airborne LiDAR Data Tools. R package version 0.10.2, https://r-lidar.github.io/lasR.

Schlager S. (2017). Morpho and Rvcg - Shape Analysis in R. En Zheng G., Li S., Szekely G. (Eds.), *Statistical Shape and Deformation Analysis*, pp. 217-256. Academic Press.

Ushey K, Allaire J, Tang Y (2024). *reticulate: Interface to 'Python'*. R package version 1.39.0, https://CRAN.R-project.org/package=reticulate.

XI Xornada de Usuarios de R en Galicia

Santiago de Compostela, 24 de outubro do 2024

R, Webscraping e Open Science. Esquivando balas en Matrix.

Álvaro Theotonio[1]

[1] Universidad Carlos III de Madrid

RESUMO

A pesares da gran cantidade de datos recollidos na actualidade e de toda a información dispoñible, en ocasións os contidos non son fácilmente accesibles para os investigadores ou a población en xeral. Neste sentido, a falta de datos en formato aberto vai necesariamente en detrimento da reproducibilidade dos estudos, da transparencia e en último termo do progreso do coñecemento científico. Ante estas limitación, os investigadores debemos de superar todos estes obsátculos para acceder á información necesaria e realizar estudos en diferentes temáticas nas que estas prácticas resultan habituais. Unha desas áreas corresponde a análise política e a composición dos órganos de goberno de cada lexislatura de diversas comunidades autónomas da nosa nación.

Máis alá das dificultades de acceso á información debido ao seu formato ou as propias características do soporte no que se atopa, a integración de diferentes paquetes de R en conxunto con extensións dos propios navegadores web permiten circunvalar todas estas trabas e crear finalmente unha base de datos axeitada. En conclusión, o presente proxecto mostra as diferentes estratexias empregadas con R para a elaboración dunha base de datos no ámbito político na que os impedimentos para o acceso á información resultaron unha tónica constante.

Palabras e frases chave: webscraping, html, extracción, vaciado, open science

AUTORES

Alonso-Pena, M. 11
Ameijeiras-Alonso, J 11
Amoedo, J.M. 13, 67
Amoroso, C. 17
Aneiros-Pérez, G. 17
Benavente-Fernández, O. 71
Blanco, A. 19
Blanco-Varela, B. 13
Bouza, C. 19
Cabana, F. 19
Campos-Romero, H. 13
Canosa-Rodrigues, A.X. 23
Cardoso-Ramalho, M.A. 42
Carrasco, I. 23
Casanova, A. 19
Crujeiras, R.M. 11
da-Silva, E.C. 42
de-Jesus-Machado, M.C. 51
de-la-Herrán, R 23
del-Río-Viqueira, I. 46
Fernández-Arias, M. 31
Fernández-Casal, R. 71
Ferreira-Alcoforado, L. 37
Flores, M. 55
Francisco-Fernández, M. 17
Fuentes-Santos, I 41
García-Martos, C. 17
Gijbels, I. 11
Ginzo-Villamayor, M.J. 59,68
Hermida, M. 23
Levy, A. 42

López-Vizcaíno, E. 46
Manuel-Amoedo, J.M. 13,67
Martínez, P. 23
Martínez-Villanueva, N. 50
Martins-dos-Santos, J.P. 51
Meira-Machado, L. 50
Navajas, R. 23
Naya-Fernández, S. 55,71
Novo-Pérez, M.A. 59,68
Obando, J.A. 63
Obando, L.N. 63
Oviedo-de-la-Fuente, M. 17
Paz-Castro, A. 67
Real, C. 19
Rincón-Ramírez, A.V. 63
Robles, F. 23
Roca-Pardiñas, J. 50
Rodríguez-Barreiro, M. 59,68
Rodríguez-Guitián, M.A. 19
Romarís-Lodeiro, N. 71
Romero, R. 19
Ruiz, C. 23
Sestelo, M. 50
Silveira-Calviño, S. 46
Tarrío-Saavedra, J. 55
Theotonio, A. 74
Vera, M. 19
Vilar-Fernández, J.A. 17

XI 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 661220