III Congreso & XIV Jornadas de Usuarios de R

6 al 8 de noviembre de 2024

Sevilla

3RqueR

Libro de resúmenes

Editado por:

Miguel Camacho Sánchez

Jerónimo Carranza Carranza

Federico Perea Rojas-Marcos

Justo Puerto Albandoz

Alberto Ramírez Moreno

Francisco Rodríguez Sánchez

Francisco Temprano García

Alberto Torrejón Valenzuela

Comité organizador

Miguel Camacho Sánchez

IFAPA SevillaR

Jerónimo Carranza Carranza

Asterionat SevillaR

Federico Perea Rojas-Marcos

Universidad de Sevilla IMUS

Justo Puerto Albandoz

Universidad de Sevilla IMUS

Alberto Ramírez Moreno

SevillaR

Francisco Rodríguez Sánchez

Universidad de Sevilla SevillaR

Francisco Temprano García

Universidad de Sevilla IMUS & SevillaR

Alberto Torrejón Valenzuela

Universidad de Sevilla IMUS & SevillaR

Comité científico

Javier Álvarez Liébana

Universidad Complutense de Madrid

Sandra Barragán Andrés

Instituto Nacional de Estadística

Jose Luis Cañadas Reche

Orange España

Inmaculada Barranco Chamorro

Universidad de Sevilla

Aurora González Vidal

Universidad de Murcia

Pedro Luis Luque Calvo

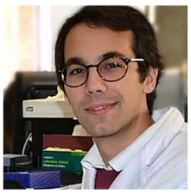
Universidad de Sevilla

Lucía Manzorro Castrillón

Instituto de Estadística y Cartografía de Andalucía

Manuel Muñoz Márquez

Universidad de Cádiz



Lluis Revilla Sancho

IrsiCaixa

Dominic Royé

Fundación para la Investigación del Clima

Cristina Tur Altarriba

Universidad de Salamanca

Iñaki Úcar

Universidad Carlos III de Madrid

Index

Programa

Miércoles 6 de noviembre			
09.00 - 09.30	Registro en el Ayuntamiento de Sevilla		
09.30 - 10.30	Inauguración en el Ayuntamiento de Sevilla		
10.30 - 12.30	Visita Archivo de Indias en grupos y plano interactivo centro de Sevilla		
12.30 - 14.30	Cóctel de Bienvenida		
15.30 - 17.30	Sesión I Premio Estudiante	Sesiones Paralelas I	Sesiones Paralelas II
17.30 - 18.30	Mesa redonda: "Acercando el dato al público general"		
18.30 - 19.15	Sesión I Póster y café en terraza IMUS		
19.15 - 20.30	Plenaria I: Rosana Ferrero		
Jueves 7 de noviembre			
09.00 - 10.00	Plenaria II: Jakub Nowosad		
10.00 - 10.30	Café en terraza IMUS		
10.30 - 12.30	Sesión II Premio Estudiante	Talleres I	Sesiones Paralelas III
12.30 - 13.30	Plenaria III: Javier Tejedor Aguilera		
13.30 - 15.30	Almuerzo en Comedor Escuela de Idiomas		
15.30 - 18.30	Sesión Premio Joven Investigador	Talleres II	Sesiones Paralelas IV
18.30 - 19.15	Sesión II Póster y café en terraza IMUS - Foto grupal		
19.15 - 20.30	Plenaria IV: Hannah Frick		
21.15 - 23.30	Cena		
Viernes 8 de noviembre			
09.00 - 12.00	Sesiones Paralelas V	Talleres III	Deliberación jurado
12.00 - 12.30	Café en terraza IMUS		
12.30 - 14.00	Asamblea, entrega de premios y clausura		

Sedes

Aula Magna Biología
Salón Actos Matemáticas
Salón Actos IMUS
Seminario IMUS

Conferencias plenarias

Plenaria 1

ROSANA FERRERO

Máxima Formación

"Más que Programar:

Cómo Formar Científicos de Datos con R"

Plenaria 2

JAKUB NOWOSAD

University of Münster,
Adam Mickiewicz University

"R's Geospatial Kaleidoscope:

Exploring Perspectives, Strengths, and Challenges"

Plenaria 3

JAVIER TEJEDOR AGUILERA

Endesa

"Retos en la Aplicación de Machine Learning en una Gran Empresa: Nuestra Experiencia"

Plenaria 4

HANNAH FRICK

Posit

"What's new with tidymodels?"

Mesa redonda

Título: *Acercando el dato al público general*

Componentes de la mesa:

- Federico Perea Rojas-Marcos *(Universidad de Sevilla)*
- Dominic Royé *(Fundación para la investigación del clima)*
- Ana Forte Deltell *(Universitat de València)*
- Juan Arévalo *(Randbee Consultants)*
- Miguel A. Armengol *(Fundación Progreso y Salud de la Junta de Andalucía)*
- Paloma López Lara *(Instituto de Estadística y Cartografía de Andalucía)*

Temas de discusión planteados:

- Democratización del dato y la alfabetización gráfica/estadística
- Democratización de la IA a través de datos FAIR
- Datos y su disponibilidad/accesibilidad
- El valor de los datos en la toma de decisiones
- La incertidumbre de los datos

Talleres y sesiones

Talleres I		
1	Inferencia Causal desmitificada	Jose Luis Cañadas-Reche
2	Machine Learning con Datos Censurados usando Tidymodels	Jesús Herranz Valera
Talleres II		
3	Taller de Visualización Analítica para la Exploración de Datos	Carlos Prieto
4	Breve introducción a la cartografía con R	Dominic Royé
5	R package dependencies in production - Risks and Management	Jan Gorecki
Talleres III		
6	Machine learning approaches for working with spatial data	Jakub Nowosad
7	Uso de R en la producción estadística: caso de uso en el IECA	Joaquin Planelles; Isabel Padilla
8	Taller de RShiny	Mireia Camacho

Sesiones de Premios		
Premio Estudiantes		
Jurado evaluador: Víctor Blanco, Gema Fernández-Avilés, Cristina Tur, Anabel Forte		
Sesión I Premio Estudiante (modera Gema Fernández-Avilés)		
1	Coexpresión y selección de variables con submuestreo aleatorio	Andrea Sánchez Moreno
2	Transcriptome Translator: A free open-source Shiny App and R Function.	Sonia María Rodríguez Huerta
3	TaphonomyR: An opensource R package for data analysis and visualization for quantitative taphonomy	Jonas Grabbe
4	Analysis of the RFSI interpolation method for the precipitation variable in Spain	Lorena Galiano Sánchez
5	Efecto de la poliploidía sobre las redes de coexpresión de genes	Alex Oliva Fernández
Sesión II Premio Estudiante (modera Víctor Blanco)		
6	LobbyBot: Análisis y Clasificación Automatizada de las Estrategias de los Grupos de Interés en las Noticias de los Medios de Comunicación de España	Aritz Gorostiza
7	Exploring class separability patterns in high-dimensional datasets with the CSVIZ tool	Marina Cuesta Santa Teresa
8	Statistical methods in R to ensure fair comparisons between treatment groups in observational studies	Natàlia Pallarès
9	Integrating R Software as a Teaching Resource for STEM Education: A Multidisciplinary Approach	Sergi Ramirez-Mitjans
10	Buscando alternativas a los índices de confort térmico en exteriores	José Antonio Rodríguez Gallego
Premio Joven Investigador		
Jurado evaluador: Víctor Blanco, Gema Fernández-Avilés, Pedro Luis Luque Calvo, Vanesa Guerrero		
Sesión Premio Joven Investigador (modera Vanesa Guerrero)		
1	CrossCarry segunda versión: Análisis de datos de un diseño cruzado mediante GEE	Nelsón A. Cruz
2	BayesianNetworks: a new tool for analysing ecological interactions	Elena Quintero Borrero
3	En busca de la eficiencia: el paquete 'labeleR' como método automático y flexible para generar etiquetas y certificados	Ignacio Ramos-Gutiérrez
4	spTReg: Modelos de regresión bayesianos espaciotemporales, el caso de la regresión cuantil	Jorge Castillo-Mateo
5	Diseño de experimentos: error absoluto vs error relativo constante	Carlos de la Calle Arroyo
6	GBS: Una aplicación "Shinyweb" para pedir cambio en las máquinas expendedoras de billetes	Tobias Kellner

Sesión Paralela I (modera Miguel Camacho)		
	1 RMoon: Una librería para cálculo astronómico	Francisco Jesús Rodríguez Aragón
	2 Predictive Modeling for Sustainable Urban Living: Machine Learning Solutions to Air Quality Challenges in Madrid	Alexandre Fabregat
	3 Maximizando el Potencial del Análisis de Comunidades con R: Herramientas Integradas para la evaluación de Artrópodos del Suelo	María del Carmen Fernández Bravo
	4 refseqR : operaciones computacionales comunes con registros de la colección RefSeq (NCBI)	Jose V. Die
	5 Identificación de OTUs diferenciales en microbiomas: Extensiones de la función explore_logratios para clasificación binaria y multinomial	Irene García Mosquera
	6 EasyAmpR: genotipado de amplicones multilocus de secuenciación masiva en R	Miguel Camacho Sánchez
Sesión Paralela II (modera Virgilio Gómez Rubio)		
	7 Moving forward in developing an innovative application for monitoring heat-related mortality in Spain	Dominic Royé
	8 Técnica de Propensity Score en estudios clínicos observacionales no aleatorizados para la estimación del efecto de un tratamiento	Irene Serrano García
	9 Análisis de supervivencia: tiempo hasta evento único y tiempo hasta eventos recurrentes	Rafael Sánchez del Hoyo
	10 MoMo: un sistema automatizado de vigilancia de la mortalidad diaria integrando R.	Inmaculada León Gómez
	11 Fuzzy Logic System for Determining Emotional and Mental States in R	Roberto Morales
	12 Estimación del exceso de defunciones durante la pandemia de COVID-19 con R	Virgilio Gómez Rubio
Sesión Paralela III (modera Francisco Rodríguez-Sánchez)		
	13 Confesión: desarrollamos software con R y Excel	Paula Cervilla; Leonardo Hansa
	14 Una Mijilla de Shiny	Álvaro Sánchez Villalba
	15 isaves: un paquete para la carga y el guardado inteligente de objetos en el workspace de R	David Hervás Marín
	16 Optimización, Solvers y R con ejemplos	Alberto Torrejón Valenzuela
	17 max_clique: the project and development of a new package for robust clustering	Stefano Benati
	18 Ensamblaje de bases de datos con validación automática utilizando GitHub Actions	Francisco Rodríguez-Sánchez
Sesión Paralela IV (modera Román Salmerón)		
	19 CRAS: Aplicación Shiny para Análisis y Simulación de Riesgos de Ciberseguridad	Emilio L. Cano
	20 Using Recurrence Analysis to search for patterns of dynamic behavior in economic time series	Lorenzo Escot
	21 Interpretación de modelos de machine learning: estimando la importancia de las variables predictoras y los efectos marginales sobre la variable objetivo	Julio E Sandubete
	22 Algoritmos basados en Generación de Columnas y Branch-and-Price	Francisco Temprano García
	23 Grafos de coincidencias y regresión	Modesto Escobar; Cristina Calvo López
	24 Selección de variables en modelos ordinales	Patricia Carracedo Garnateo
	25 Estimación cresta generalizada en R para el modelo de regresión lineal múltiple	Román Salmerón Gómez
Sesión Paralela V (modera Kamal Antonio Romero)		
	26 R en la lucha contra los incendios forestales: conociendo la evolución de un incendio.	Marta Rodríguez Barreiro
	27 Incidencia delictiva en municipios españoles: análisis descriptivo y factores determinantes	Pedro Jose Perez Vazquez
	28 El efecto de factores socioeconómicos sobre la conectividad del transporte público	Javier Hernán Matas Monroy
	29 Una aplicación de datos abiertos en la Administración pública: El Panel de Indicadores de Turismo de Lanzarote	José M. Cazorla Artiles
	30 Utilizando R para el análisis de tablas de tres entradas: una aplicación en datos sobre internacionalización en la educación terciaria	Álvaro Toledo San Martín
	31 Development of our own R packages for the statistical processes of the Generic Statistical Business Process Model	Elisa Jorge
	32 Fundamentos de ciencia de datos con R	Gema Fernández-Avilés Calderón
	33 Cuentas nacionales a gran escala con R: FIGARO multi-country input output tables	Kamal Antonio Romero Sookoo

Sesión I Pósters		
	1 Una app Shiny para la visualización de los resultados del Registro OHSCAR	Patricia Fernández del Valle
	2 Nuevos retos en los modelos desarrollados en R para la evaluación y gestión de pesquerías bajo incertidumbre	María Soto Ruiz
	3 phyloshapeR: plotting phylogenies to look like maps	Ignacio Ramos-Gutiérrez
	4 R en la Evaluación de Estrategias de Gestión de Stocks de Recursos Marinos	Diana María González Troncoso
	5 Igualdad de Género en la Transferencia de Conocimiento: Análisis Avanzado con R	Matilde Pulido Prior
	6 La implicación de la mujer en la innovación y la transferencia de conocimiento a través del análisis de la generación de patentes	María Isabel Sánchez-Rodríguez
Sesión II Pósters		
	7 Evaluación de RSPRITE en la Detección de Fraude en Ciencias Sociales"	Antonio Matas-Terrón
	8 lp.edu: un paquete para introducir las técnicas de optimización lineal en estudios universitarios de grado	Josep Antoni Martin Fernandez
	9 Desarrollo de un paquete en R para la implementación de operadores SUOWA y Semi-SUOWA	Teresa Gonzalez Arteaga
	10 Seguridad para las brigadas terrestres de extinción de incendios: cálculo con R de la ruta más rápida a un lugar seguro.	Manuel Antonio Novo Pérez
	11 Desarrollo de un sistema de captura y análisis de datos de ensayo clínico en el ámbito hospitalario basado en herramientas de libre acceso	Mateo Paz Cabezas

Contribuciones

Talleres

Inferencia Causal desmitificada

Autor: Jose Luis Cañadas-Reche [1]

[1] *Orange Spain*

No es tan díficil ni tan oscuro como piensan algunos. En realidad, ya sabes inferencia causal.
En este taller trataré de explicar algunos conceptos, no están todos los que son, y puede que algunos no se den. Y puede que tampoco en este orden.

- Ciencia antes que estadística
- Inferencia causal no es más que predecir el efecto de la intervención
- DAG's.
- Algunas reglas de Pearl
- Saltarnos reglas de Pearl gracias a que somos Bayesianos. Full luxury
- Otras técnicas: clásicas y modernas (son sólo técnicas, lo importante es lo de antes)
- Propensity score matching y relacionados
- Metalearners
- Doubly robust estimation
- Y recordad, si podéis hacer un experimento bien diseñado, mucho mejor. Esto de la inferencia causal es sólo un intento de poder sacar cosas útiles cuando no podemos hacer un experimento.

Agradecimiento especial a Richard McElreath

Machine Learning con Datos Censurados usando Tidymodels

Autor: Jesús Herranz Valera [1]

[1] *Fundación Grupo Español de Investigación en Cáncer de Mama (GEICAM)*

Tidymodels es un metapaquete, donde se han integrado todos los procesos de construcción y evaluación de modelos predictivos, manteniendo la filosofía de programación de tidyverse. Centrado fundamentalmente en problemas de regresión y clasificación, tiene extensiones que permiten trabajar con datos censurados, datos de supervivencia.
El uso de datos censurados cada vez se está extendiendo más, como en problemas relacionados con la fidelización de productos y clientes en las empresas, y por otro lado, es fundamental en algunas áreas de investigación biomédica, como son la oncología y las enfermedades cardiovasculares
Este taller abarcará todas las fases de la construcción de un modelo predictivo con el paquete tidymodels y sus extensiones, con un ejemplo de datos de supervivencia de alta dimensionalidad (p »n). Se explicarán las etapas de pre-procesamiento de datos con "recipe", medidas de la capacidad predictiva, como son el c-index, el brier score o los AUCs de las time-dependent ROC curves, con "yardstick". La construcción de los modelos de supervivencia con el paquete "censored", que es la extensión de "parsnip" para este tipo de datos. También se explicará la evaluación y optimización de los parámetros del modelo con muestras de training y testing, y con técnicas de remuestreo implementadas en los paquetes "rsample" y "tune".
Se explicarán desde modelos de supervivencia básicos, como es la regresión de Cox, hasta modelos de machine learning con datos censurados, como son Random Survival Forest o Boosting. Además,

se establecerán comparaciones entre el rendimiento predictivo de estos modelos, integrándolos en un "workflow", una de las facilidades fundamentales que permite tidymodels

Taller de Visualización Analítica para la Exploración de Datos

Autor: Carlos Prieto [1]

Co-autores: Modesto Escobar [1]; David Barrios [1]

[1] *Universidad de Salamanca.*

En los últimos años, el desarrollo de tecnologías Web Front-end ha impulsado la creación de nuevas herramientas de visualización en R. Estas herramientas permiten una visualización interactiva y dinámica de datos en un navegador web, facilitando la exploración y análisis de resultados obtenidos mediante diversas técnicas. En este taller, aprenderemos sobre nuevas soluciones de visualización que permiten:
• La interpretación analítica de resultados
• Realizar una exploración visual e interactiva los datos.
• Mostrar la información de bases de datos en aplicaciones web multimedia.
• Realizar Data Storytelling de proyectos.

Objetivos
• Explorar las posibilidades de las representaciones interactivas.
• Crear visualizaciones interactivas, dinámicas y analíticas.

Agenda
1. Visualización Interactiva para la Descripción de Datos: Utilizaremos el paquete RJSplot para describir y explorar datos de forma interactiva.
2. Herramientas para la Visualización de Pruebas de Contraste, Técnicas de Clustering y Reducción Dimensional: Aplicaremos Rvisdiff para visualizar los resultados de pruebas de contraste. Utilizaremos looking4clusters para generar y explorar resultados de técnicas reducción dimensional y clustering.
3. Creación de Mapas Interactivos y Evolutivos en el Tiempo: Mediante el paquete evolMap representaremos la información de una base de datos sobre un mapa geográfico interactivo. Los datos se representarán sobre el mapa mediante marcadores, líneas o coropletas, que pueden adaptar su aspecto visual en función de la información de la base de datos, y mostrar su evolución en el tiempo.
4. Creación de Redes Analíticas y Dinámicas: Emplearemos el paquete netCoin para generar herramientas de visualización analítica que representan redes dinámicas conectadas a tablas de información.
5. Generación de una Web Multimedia desde una Base de Datos: Con las funciones gallery2 y netGallery2 de netCoin, crearemos una web multimedia que muestre una galería fotográfica vinculada a información detallada, facilitando su estudio y exploración.

Conclusión
Al finalizar este taller, los participantes habrán adquirido habilidades para crear y utilizar herramientas de visualización analítica avanzadas, mejorando significativamente su capacidad para explorar y presentar datos de manera interactiva y dinámica.

Financiación
D.B. ha sido financiado por el programa PTA (PTA2022-022270-I) del Ministerio de Ciencia, Innovación y Universidades. La elaboración de los paquetes ha sido financiada por los siguientes proyectos: Analytic Networks for Dissemination and Research (PDC2022-133355-I00) y Network Coincidence Analysis (PGC2018-093755-B-I00), ambos financiados por el Ministerio de Ciencia, Innovación y Universidades.

Breve introducción a la cartografía con R

Autor: Dominic Royé [1]

[1] *Fundación para la Investigación del Clima (FIC)*

Podemos encontrar datos espacio-temporales en cualquier lugar. Nos encontramos con información espacial en cualquier aspecto de nuestra vida cotidiana desde la televisión, los periódicos, la informática, los móviles o directamente en los mapas. Creamos información espacial con muchas de nuestras actividades, especialmente con el uso masivo de tecnología desde dispositivos móviles. Sin embargo, para obtener una visualización adecuada, cada vez es más importante hacer uso de programación. En este taller veremos brevemente las bases sobre el uso de datos vectoriales y ráster para después poder visualizarlos mediante el paquete ggplot2 apoyándose en otros paquetes auxiliares. Desde un mapa de densidad de puntos, sobre un mapa clásico de coropletas hasta un cartograma.

R package dependencies in production - Risks and Management

Autor: Jan Gorecki [1]

[1] *Freelance*

R packages have a very convenient mechanism for stating their R dependencies, and then resolving them during installation.
This convenience resulted in developers no longer worrying about adding extra dependencies to their packages.
We will briefly look at dependency relationships between R packages, as well as OS dependencies of R packages. We will summarize what is the current state of dependencies usage in R ecosystem. Then present dependencies from the other angle by identifying risks that are being introduced when a package adds dependencies. Ultimately we will discuss best practices and provide examples of simple changes that will reduce risks introduced by extra dependencies, which anyone deploying R in production should consider.

Machine learning approaches for working with spatial data

Autor: Jakub Nowosad [1]

[1] *University of Münster, Adam Mickiewicz University*

The 'Machine Learning Approaches for Working with Spatial Data' workshop highlights the similarities and differences between machine learning using spatial data and non-spatial data. The workshop guides participants through various stages of machine learning workflows, from data preparation to model evaluation and prediction. A traditional machine learning workflow will be discussed, followed by specific approaches for dealing with spatial data. These include spatial feature engineering, spatial cross-validation, area of applicability, and model explainability. The workshop will use reproducible code, plots, and flowcharts to illustrate spatial machine learning workflows and methodologies.

More: https://github.com/Nowosad/IIIRqueR_workshop_materials

Uso de R en producción estadística: caso de uso en el IECA

Autor: Joaquin Planelles[1]

Co-autor: Isabel Padilla Sánchez [1]

[1] *Instituto de Estadística y Cartografía de Andalucía*

En cualquier trabajo con datos hay distintas fases. Típicamente el proceso comienza con la lectura/captación de la información. Y esto puede implicar acciones muy diversas. En unos casos serás tú mismo el que capte esa información (preguntando, observando, sensorizando) y en otras ocasiones -las más- reutilizarás información que ha sido captada por terceros con fines distintos a la producción de estadísticas o modelización de los datos. Por ejemplo, información que ha ido quedando sedimentada en una base de datos para la prestación de un servicio. En otras ocasiones acudirás a información menos estructurada, haciendo Web-scrapping o tratando imágenes (aquí la escala va de los satélites a los cultivos microscópicos en laboratorio).

Una vez has captado la información, el siguiente paso es la transformación de la estructura y contenido de tu set de datos. Esta segunda fase, dedicada al tratamiento de datos, es la "fase gris" del ciclo de vida del dato. Le falta el atractivo visual que tienen las fases siguientes, centradas en la modelización y la comunicación. No obstante, por experiencia sabemos que una parte significativa del tiempo total de trabajo se dedica a este tipo de operaciones. Esto es así en cualquier análisis de datos, pero en mayor medida en el caso de la producción estadística, en la que podemos afirmar que la mayor parte del trabajo consiste precisamente en este tipo de operaciones.

Siendo esto así, existen distintas herramientas para abordar el tratamiento de datos. Una de ellas, que en nuestra organización está cada vez más presente, son precisamente un conjunto de librerías de R/tidyverse que generan scripts sencillos, permiten industrializar e integrar los procesos, así como obtener una respuesta eficiente en tiempos de procesamiento.

El objetivo de este taller es facilitar algunos scripts que a nosotros nos han resultado útiles a la hora de acceder a información, y también para transformarla. Se trata de un muestrario de soluciones que nosotros hemos implementado en algún proyecto y que pensamos que, aunque sea en ámbitos temáticos distintos, o como fase previa a la modelización, a otros usuarios también les pueden resultar útiles.

Taller de RShiny

Autor: Mireia Camacho [1]

[1] *Universidad Autónoma de Barcelona*

Explicación paso a paso de la elaboración de una aplicación desarrollada con RShiny. El objetivo sería explicar las posibilidades que ofrece (interactividad, dashboards, visualizaciones, puesta de modelos en producción…), la lógica que sigue el código, buenas prácticas, trucos para facilitar el desarrollo…a la vez que iremos mostrando paso a paso el desarrollo de una aplicación a lo largo del taller. También se incluirá la subida de la app a un servidor Shinyapps y se explicaran otras alternativas de puesta en producción.

Contribuciones orales

RMoon: Una librería para cálculo astronómico

Autor: Francisco Jesús Rodríguez Aragón [1]

[1] *Universidad de Córdoba; Oney Servicios Financieros*

La Luna es un satélite de un tamaño excesivo con respecto al planeta que orbita, esto hace que sea especialmente difícil la predicción certera de elementos que la mayoría del público supone esenciales, de hecho en la misión Apolo de 1969 Amstrong tuvo que tomar el control de la nave porque se estaba desviando del punto de aterrizaje. Así pues, elementos como son la posición de la Luna en el cielo en una fecha determinada, cuándo va a tener exactamente lugar un eclipse lunar o solar y cuáles van a ser sus características esenciales, qué elementos clave de su superficie serán o no observables, cuándo tiene lugar un perigeo o un apogeo y a qué distancia, etc, son problemas interesantes para los que se requiere algoritmos que tengan en cuenta muchas perturbaciones y efectos gravitatorios y geométricos.

En este sentido, se crea la librería RMoon que va creciendo conforme avanza el tiempo y que se encuentra disponible en github: https://github.com/FJROAR/RMoon

Esta librería ofrece un amplio repertorio de funciones y datos, muchos de ellos basados en el libro Astronomical Algorithm de Jean Meeus que permiten resolver muchas de las anteriores cuestiones, de modo vectorizado y bajo un enfoque puramente R, con un grado de exactitud bastante aceptable, por lo que en este taller se pretende mostrar al interesado, una introducción práctica a los principales problemas que plantea el estudio de nuestro satélite a través de algunas actividades y ejemplos que se acompañarán para facilitar la comprensión de los principales conceptos clave referentes a nuestra Luna.

isaves: un paquete para la carga y el guardado inteligente de objetos en el workspace de R

Autor: David Hervás Marín [1]

Co-autor: Patricia Carracedo Garnateo [1]

[1] *Universitat Politècnica de València*

Cuando se trabaja con proyectos de análisis de datos complejos es usual almacenar numerosos objetos en el workspace de R. Si además de tener numerosos objetos estos son de un tamaño considerable, el manejo de dicho workspace puede volverse engorroso a la vez que ineficiente computacionalmente.
Para mejorar el flujo de trabajo en estas situaciones hemos creado el paquete de R isaves (disponible en github). Este paquete permite agilizar los procesos de guardado y de carga de objetos en el espacio de trabajo proporcionando, entre otras, las siguientes funcionalidades: guardado y carga incremental del espacio de trabajo, guardado y carga selectiva de objetos en el espacio de trabajo, lazy loading, historial de versiones de objetos con posibilidad de recuperar versiones antiguas de los mismos, base de datos con metadatos de cada objeto, etc.
En este trabajo presentamos la funcionalidad completa del paquete isaves, así como un ejemplo de su aplicación para trabajar en un proyecto de análisis de datos complejo.

Predictive Modeling for Sustainable Urban Living: Machine Learning Solutions to Air Quality Challenges in Madrid

Autor: Alexandre Fabregat [1]

Co-autor: Anton Vernet [1]

[1] *Universitat Rovira i Virgili*

Accurate prediction of air quality in urban environments is key to designing public health interventions aimed at reducing exposure to harmful air pollutants.
Using the case of Madrid, here we present a comprehensive machine learning approach to better understand how different factors affect local air quality across the metropolitan area.
Reducing pollution levels is a pressing need for this city that often fails to meet the standards set by the European Commission. To elucidate the relationship between pollutant concentration (the response) and the factors that we hypothesize affect it most, namely local atmospheric conditions and road traffic (the predictors), we processed 24 months of hourly data on the number of vehicles and meteorology together with the concentration of nitrogen oxides, ozone and particulate matter measured at 22 stations spread across the metropolitan area. We used the R language to preprocess the data and train a model for each pollutant station and chemical species following the workflow of the "tidymodels" framework. In terms of predictive capability, the present results show at least comparable performance to other existing models. This is especially remarkable given that the present model uses a parsimonious collection of predictors, all of them coming from public and freely available sources. In terms of computational cost, the present models require far fewer resources than the state-of-the-art pollutant dispersion models that solve some form of differential equations governing the transport of chemical species in the turbulent planetary boundary layer. The present work can be used directly to evaluate the effectiveness of various measures and strategies aimed at reducing the level of pollutants, such as, for example, restricting road traffic access to some city districts. Overall, this research helps to inform urban developers, city planners, and public health officials in their mission to achieve cleaner and healthier cities.

Selección de variables en modelos ordinales

Autor: Patricia Carracedo Garnateo [1]

Co-autores: David Hervás Marín [1]; Raquel Soriano [1]

[1] *Universitat Politècnica de València*

Los modelos predictivos con respuesta ordinal son herramientas fundamentales en el análisis de datos cuando la variable de interés es ordinal. En concreto, en el ámbito actuarial, esta situación es muy común puesto que variables que miden la calidad crediticia del cliente mediante escalas. La escasez de estudios sobre estos modelos es una realidad, lo que impide una comprensión más profunda y precisa de los datos. Este trabajo propone varios métodos para seleccionar variables de interés en los modelos estadísticos como elastic net, stepwise, lasso entre otras. Además, se detallaran las librerías utilizadas.

Exploring class separability patterns in high-dimensional datasets with the CSVIZ tool

Autor: Marina Cuesta Santa Teresa [1]

Co-autores: Alberto Fernández-Isabel [1]; Carmen Lancho [1]; Emilio López Cano [1]; Isaac M. de Diego [1]

[1] *Universidad Rey Juan Carlos*

In a Data Science project, data visualization stands as an essential asset in several phases of its lifecycle. In particular, it is crucial during the Exploratory Data Analysis, supporting the discovery of

anomalies, structure, and relationships within the data to fully understand the underlying problems. In classification problems, data visualization is helpful to reveal class separability patterns within the dataset by visually exploring the class distributions and topology. It enables the identification of regions within the feature space where classes are well-separated or overlapped. This is very valuable information that can be later used when building a Machine Learning model, helping to choose an appropriate one and to improve the model performance.

In this context, high-dimensional data arise as a challenge. Traditional visualization techniques such as the scatterplot matrix struggle with their representation due to space limitations to correctly display them. With d variables in a dataset, the scatterplot matrix must include $\frac{d \cdot (d-1)}{2}$ pairwise scatterplots in a single display. As d increases, the graph becomes overplotted. Additionally, humans lack the cognitive ability to discover structures and patterns within a huge scatterplot matrix. Alternative visualization methods often rely on dimensionality reduction techniques, yet this can compromise interpretability as patterns are explored in a transformed space of the original variables. Indeed, in many applications, it's critical to keep the original variables rather than a transformation of them to make informed decisions.

Acknowledging the previously discussed issue, the authors have proposed the Class Separability Visualization (CSViz) method as a new Visual Analytics tool to deal with the visualization of labeled high-dimensional data in their original variables. CSViz addresses this challenge with a subspace approach. It offers a set of 2-Dimensional subspace visualizations, each containing exclusive subsets of points from the original variables with the most valuable and significant separable patterns within the dataset. Thus, CSViz offers an overview of the class separability in the dataset, reducing the number of scatterplots to be inspected compared to the scatterplot matrix. CSViz significantly eases the visual exploration in the EDA, and thus, reduces the amount of time invested in it.

The CSViz method has been implemented using the R software, and the open-source code can be found at https://github.com/URJCDSLab/CSViz.

Confesión: desarrollamos software con R y Excel

Autor: Leonardo Hansa [1]

Co-autor: Paula Cervilla [1]

[1] *Ebiquity*

Trabajamos en una empresa que lleva estudiando la eficacia publicitaria con modelos econométricos un mogollón de años. En el pasado lo hacían con EViews pero llegó un momento en que se animaron a dar el salto al software libre. **Se decantaron por R.**

Quizá por aquella época Python aún no estaba tan extendido a datos, o quizá simplemente no se habían enterado de su existencia, el caso es que parece que **el eslogan ese de "R no sirve para producción" no cuajó en Ebiquity**, y gracias a eso R se instauró como el lenguaje de referencia.

El inconveniente de R era que, dado que la mayoría de miembros del equipo no programaba, la curva de aprendizaje sería muy elevada. Y el objetivo no era que se convirtieran en programadores, sino proveerlos de las mejores herramientas para que hicieran sus modelos econométricos cómodamente. Nos gustara o no, **Excel jugaría un papel importante.**

La solución fue el ecosistema *ebverse*, un conjunto de librerías de R desarrolladas internamente, que sirve como backend de una herramienta basada en Excel. Esta herramienta hace modelos econométricos con datos de panel.

Es una plantilla de Excel tuneada, con un aspecto que resultará familiar a cualquier usuario de Excel. Y además tiene botoncitos que llaman a **VBA** para a su vez llamar a R. Todo el análisis estadístico y preparación de datos se hace con R, aunque el usuario no tiene por qué saber usarlo (salvo un poco de sintaxis para especificar las variables que quiere en el modelo).

La plantilla de Excel es muy cuadriculada, y si bien el desarrollo de software en R nos mola, en Excel es un horror. Así que las nuevas funcionalidades se hacen en **Shiny.**

Aparte de nuestras librerías de procesamiento en R, tenemos nuestras librerías en Shiny. Seguimos un framework parecido a Golem con el que desarrollamos shiny apps como librerías de R, lo que nos facilita la instalación para todos los usuarios. Los shiny nos sirven para funcionalidades para las que Excel se nos queda corto.

En nuestra comunicación **presentaremos** algunas de nuestras **librerías favoritas** del ecosistema, así como el **objetivo de algunas Shiny** apps y el papel de **VBA** en todo esto.

TaphonomyR: An open-source R package for data analysis and visualization for quantitative taphonomy

Autor: Jonas Grabbe [1]

Co-autores: Antonio Canepa-Oneto [1]; Ana Serrano-Mamolar [1]; Ana Pantoja-Pérez [2]; César García-Osorio [1]; Nohemi Sala [2]

[1] *Universidad de Burgos*

[2] *Centro Nacional de Investigación sobre la Evolución Humana (CENIEH)*

In the fields of paleontology and archaeology, quantitative taphonomy plays an important role in deciphering the interplay between different agents including human activities, carnivores and other natural processes on skeletal remains.

Despite the critical insights it offers, the field faces challenges including a lack of standardized data preparation, protocols and difficulties in reproducibility of results.

By combining data science techniques with taphonomic theory, our goal is to provide a tool that will aid in the understanding of taphonomic processes and improve and streamline the analysis of hominin and faunal remains. "TaphonomyR" facilitates data standardization and exploration, modeling and visualization for skeletal part abundance, bone density and breakage patterns data.

The strength of the 'TaphonomyR' package lies in its integrated dataset of archaeological sites, which facilitates comparative studies and enhances our understanding of new findings in relation to established categories of archaeological sites, particularly through its accompanying user-friendly Shiny App.

"TaphonomyR" not only advances archaeological research but also fosters interdisciplinary collaboration, setting a new standard for reproducibility, methodological rigor, and open science in archaeology and paleontology.

Una Mijilla de Shiny

Autor: Álvaro Sánchez Villalba [1]

[1] *Appsilon*

Lecciones aprendidas, mejores prácticas y experiencias acumuladas a través de mi experiencia creando aplicaciones web con R Shiny para empresas en el Fortune 500.
En esta charla, compartiré una «mijilla» de conocimientos que os ayudarán a mejorar varios aspectos de vuestras aplicaciones Shiny: rendimiento, interfaces de usuario, accesibilidad, código, tests, etc.

Estimación cresta generalizada en R para el modelo de regresión lineal múltiple

Autor: Román Salmerón Gómez [1]

Co-autores: Catalina García García [1]; Guillermo Hortal Reina [1]

[1] *Universidad de Granada*

La regresión cresta (regular) es posiblemente la técnica de estimación alternativa a los mínimos cuadrados ordinarios (MCO) más usada para ajustar un modelo de regresión lineal múltiple cuando en éste existe un problema de relacionales lineales (multicolinealidad) preocupante. Esta técnica se caracteriza por proporcionar estimadores sesgados con menor error cuadrático medio (ECM) que los estimadores por MCO. En cambio, dentro de nuestro conocimiento, su versión generalizada no ha sido ni desarrollada teóricamente en profundidad ni aplicada.
En el trabajo propuesto se obtiene una expresión cerrada para su estimador, norma, error cuadrático medio (ECM), bondad de ajuste y se propone hacer inferencia bootstrap. Igualmente se aplica buscando obtener estimadores (sesgados) con menor ECM que los estimadores proporcionados por MCO y la regresión cresta regular.
Finalmente, también se comparan los resultados obtenidos con los proporcionados con paquetes ya existentes en R para la obtención del estimador cresta regular.

Técnica de Propensity Score en estudios clínicos observacionales no aleatorizados para la estimación del efecto de un tratamiento

Autor: Irene Serrano García [1]

Co-autor: Rafael Sánchez del Hoyo [1]

[1] *Unidad de Apoyo Metodológico a la Investigación, Hospital Clínico San Carlos (IdISSC)*

En investigación clínica, el mejor diseño de estudio metodológicamente es el ensayo clínico controlado y aleatorizado. Cuando se quiere comparar dos tratamientos, o dos técnicas (dos grupos), cada paciente recibirá uno u otro en función de una decisión aleatoria. Por tanto, se reduce e incluso se eliminan sesgos que surgen de las características de los pacientes. De esta forma, se permite estimar los efectos del tratamiento o técnica comparando directamente los resultados entre los pacientes de un grupo u otro.

Pero no siempre es posible realizar un ensayo clínico (por costes, participación, etc.), y se realizan estudios observacionales, donde se analiza la práctica clínica habitual. A los pacientes se les observa, se recogen los datos y se comparan. El inconveniente es que puede haber diferencias sistemáticas en las características basales de los pacientes entre un grupo y otro. Esto confunde el resultado del efecto del tratamiento o técnica.

Para poder corregir las diferencias basales entre los grupos, existen diferentes técnicas como el emparejamiento. A través de la técnica de Propensity Score (PS) se van a crear estos emparejamientos.

El proceso de análisis en los estudios observacionales sería el siguiente. Primero se analizan las diferencias entre las características basales de los grupos. Si se encuentran diferencias estadísticamente significativas al 95% de confianza, se procede a estimar el PS. Este score se calcula mediante las variables basales que interesan que sean homogéneas para que no aporten sesgo, y el score es la probabilidad predicha de pertenecer a uno de los grupos. Una vez se tiene esta puntuación, se realiza el emparejamiento y se evalúa la homogeneidad de las variables implicadas. Si el PS es adecuado, finalmente se procede a estimar el efecto del tratamiento o técnica.

Este emparejamiento se puede ejecutar a través de 3 técnicas diferente: el *vecino más cercano*, *optimal matching*, o *genetic matching*.

En nuestra Unidad se ha realizado un trabajo donde se quiso analizar el efecto (en términos de menor infección) con el uso de un protector en herida quirúrgica para el tratamiento de apendicitis aguda. Al analizar los dos grupos de estudio: con protector y sin protector; se observaron diferencias estadísticamente significativas en ciertas características basales. Se consideraron que 4 de ellas eran relevantes para sesgar el efecto del protector; por tanto, se realizó un PS con estas 4 variables.

La librería que se usó con R es **MatchIT**, que a través de la función **matchit** permite crear los pares con una puntuación similar. De esta manera, la muestra emparejada, es la que tiene las variables homogeneizadas y así poder estimar el efecto de este protector sin sesgos iniciales.

Con esta función se pueden realizar las 3 técnicas diferentes del PS. A través del argumento *method* se puede elegir entre *nearest*, *optimal* o *genetic* y crear los distintos emparejamiento.

Otros de los argumentos importantes es el ratio. Para crear parejas, se escogió un ratio 1:1, pero en función del tamaño de los grupos, se puede modificar y escoger 1:2 para tener, por ejemplo, un paciente con protector y 2 sin protector. El otro que hay que explorar es el *caliper* , que es una restricción en la distancia entre las puntuaciones para formar parejas. Así, si tenemos algún paciente que solo se pueda emparejar con una distancia mayor que el *caliper*, por tanto, una diferencia de PS alta, se quedaría sin emparejamiento asegurando la homogeneización de las variables basales.

Palabras clave: Propensity Score, Emparejamiento, Estudio Observacional MatchIt

Statistical methods in R to ensure fair comparisons between treatment groups in observational studies

Autor: Natàlia Pallarès [1]

Co-autor: Cristian Tebé [1]

[1] *Institut de Recerca Germans Trias i Pujol (IGTP)*

Randomised clinical trials (RCTs) are considered the gold standard for studying the effectiveness of interventions or treatments because randomization ensures similar baseline characteristics and eliminates confounding variables. Observational studies do not use randomization, leading to differences between groups in measured or unmeasured characteristics that could confound the association between the exposure and the outcome being studied.

Several statistical methods have been developed to control for confounding in observational studies. Multivariable regression is the traditional method and propensity score (PS) methods are an alternative. The range of methods that use the PS to correct for baseline differences between groups includes using the PS as an adjustment covariate in a regression model, propensity score matching (PSM), and inverse probability weighting (IPW).

When examining differences in mortality between waves of COVID in Catalonia, we need to make the waves comparable in terms of baseline covariates to ensure fair comparisons. We will illustrate how these methods can be applied in this scenario using different R packages: MatchIt and WeightIt to apply PSM and IPW; cobalt to check covariate balance after matching or weighting; survey to fit models with weights obtained from IPW, and forestplot to graphically compare the results of the four techniques.

We will compare the results of the four methods, present arguments for and against each one of them, and make recommendations for comparing treatment groups in observational studies.

Análisis de supervivencia: tiempo hasta evento único y tiempo hasta eventos recurrentes

Autor: Rafael Sánchez del Hoyo [1]

Co-autor: Irene Serrano García [1]

[1] *Unidad de Apoyo Metodológico a la Investigación; Hospital Clínico San Carlos; (IdISSC)*

En investigación clínica son muy comunes los estudios longitudinales, que se caracterizan por seguir a los pacientes durante un período de tiempo el cual puedes ser variable o presentar observaciones incompletas. Debido a esto, surge la necesidad de emplear una técnica estadística que no solo analice la aparición de un evento (fallecimiento, recaída) sino también es necesario saber el tiempo que transcurre hasta aparecer ese evento. Para ello, existe el **análisis de supervivencia** que es una técnica que analiza respuestas binarias en estudios longitudinales.

El análisis de supervivencia no solo tiene en cuenta el tiempo hasta que ocurra el evento, sino también hasta que no ocurra (fin período de observación), o que simplemente, en cierto momento, se ha detenido el seguimiento del paciente (a lo último se le denominan valores censurados).

En nuestra unidad de apoyo a la investigación, además de analizar la mediana de seguimiento de una población en concreto, realizamos estudios oncológicos, cardiológicos y de otros ámbitos clínicos, en los que se observa el efecto de seguir distintas líneas de tratamiento o de tener ciertas características basales diferentes. Este efecto se estima a raíz del **Hazard Ratio** que se calcula mediante la **regresión de Cox**. Estos análisis los realizamos en R empleando las funciones "survfit" y "coxph" de la librería "survival".

La representación gráfica de las curvas de supervivencia se realiza con las gráficas de **Kaplan-Meier**. Para ello se utiliza el paquete "ggsurvplot" de la librería "survminer".

A diferencia del evento exitus, hay otro tipo de eventos como el empeoramiento o la recaída que pueden ocurrir más de una vez para cada sujeto a lo largo del periodo de observación. Para ello, se emplea el **análisis de supervivencia con eventos recurrentes**. Utilizamos las mismas librerías y funciones de R que en el análisis de supervivencia, pero en este caso se utilizan argumentos distintos. Además, el formato de la base de datos varía.

Estas técnicas no solo son utilizadas en el ámbito sanitario, también pueden aplicarse en otros campos como la fidelización de clientes, el tiempo hasta la obsolescencia de un producto o la migración de empleados.

Palabras claves: Supervivencia, eventos recurrentes, bioestadística, estudio longitudinal y survfit.

Incidencia delictiva en municipios españoles: análisis descriptivo y factores determinantes

Autor: Pedro Jose Perez Vazquez [1]

[1] *Universitat de València*

La ponencia aborda el análisis de la delincuencia en España mediante el empleo de datos obtenidos de Eurostat, INE y el Ministerio del Interior. Se emplea un enfoque metodológico que combina técnicas estadísticas y herramientas de análisis de datos, con énfasis en el uso del software R. El uso de R como herramienta analítica permite realizar un análisis exhaustivo de los datos, incluyendo técnicas de visualización, modelado estadístico y pruebas de hipótesis.
El objetivo principal del trabajo es examinar si variables socioeconómicas, como la renta y la desigualdad, influyen en la tasa de criminalidad. Utilizando un enfoque empírico, se lleva a cabo un análisis detallado que revela patrones significativos en la relación entre estos factores y la incidencia delictiva.

En una primera parte del trabajo se utilizan datos de Eurostat para contextualizar y poner en dimensión la verdadera dimensión de la delincuencia en España: los datos indican que España presenta, en general, unas menores tasas de criminalidad que los de los principales países de su entorno. En una segunda parte se utilizan datos de criminalidad provenientes del Ministerio del Interior para obtener las principales pautas de la criminalidad en las diferentes provincias españolas. Señalar que los datos del Ministerio presentan información sobre 13 tipos de delitos distintos.
En una tercera sección, se utilizan datos de criminalidad del ministerio pero a nivel municipal para los municipios con población superior a 20.000 habitantes, para realizar un análisis exploratorio y descriptivo sobre la incidencia de los diversos delitos a nivel municipal. Los resultados indican que, el tamaño del municipio y la presencia de costa (seguramente ligado a las actividades turísticas) son importantes factores para explicar las diferencias en las incidencia delictiva. El análisis también detecta la existencia de áreas o clusters de municipios cercanos con mayor incidencia de la delincuencia.
Por último, en la cuarta sección, se sigue utilizando los datos de criminalidad del ministerio a nivel municipal y se recopila información socioeconómica relevante para, mediante el uso de modelos estadísticos, examinar si variables socioeconómicas, como la renta y la desigualdad, influyen en la tasa de criminalidad a nivel municipal. Para ello se utilizan diversos indicadores sobre pobreza, desigualdad, tasa de paro y de la población (porcentaje de población joven y población extranjera), todos provenientes de diversas operaciones estadísticas del INE, como por ejemplo, el Atlas Experimental de distribución de la Renta.
Los resultados indican una asociación clara entre niveles más bajos de renta y un aumento en la tasa de criminalidad, sugiriendo que la privación económica puede ser un factor motivador importante para la participación en actividades delictivas. Asimismo, se encuentra evidencia de que la desigualdad económica también puede contribuir al aumento de la criminalidad, ya que áreas con mayores disparidades de ingresos tienden a exhibir tasas de criminalidad más altas.
Este análisis contribuye a la comprensión de los determinantes subyacentes de la delincuencia en España, ofreciendo información valiosa para la formulación de políticas públicas orientadas a la prevención y el control del crimen.
En conclusión, el análisis que combina diversas fuentes de datos con técnicas avanzadas de análisis de datos, respaldadas por el uso de R, contribuye a la comprensión de los determinantes subyacentes de la delincuencia en España aportando evidencia empírica sólida sobre la relación entre variables económicas y la tasa de criminalidad en España.

Desarrollo de un sistema de captura y análisis de datos de ensayo clínico en el ámbito hospitalario basado en herramientas de libre acceso

Autor: Mateo Paz Cabezas [1]

[1] *Unidad de Apoyo Metodológico a la Investigación; Hospital Clínico San Carlos; (IdISSC)*

Durante el proceso del desarrollo del ensayo clínico se generan y almacenan una gran cantidad de datos. Debido a la rigidez de los sistemas hospitalarios, la sensibilidad de los datos recogidos y a la heterogeneidad organizativa entre los distintos servicios de Oncología Médica, dichos datos se almacenan en la mayor parte de los casos de forma no estructurada, atomizada en distintos archivos y soportes y sin una estructura que permita su análisis a gran escala.
Material y métodos
El sistema de captura de datos RedCap nos permite desarrollar instrumentos para el registro de información de propuestas, contratos, desarrollo logístico y clínico de los ensayos, manteniendo los datos en la intranet del propio centro, evitando conflictos con la propiedad o confidencialidad de los mismos. Con RStudio, automatizamos la descarga de los datos. Finalmente, el software PowerBI nos permite analizar y visualizar los datos en cuadros de mando. Al tratarse de herramientas de libre acceso para centros de investigación, no repercute coste alguno para el equipo investigador.
Resultados
El análisis en tiempo real de la información permite orientar la actividad investigadora a la obtención de resultados, definiendo KPIs (key performance indicators) personalizados que nos permitan detectar debilidades y diseñar estrategias de mejora para cada uno de los procesos implicados en el desarrollo de los ensayos, aumentando los beneficios obtenidos a nivel clínico y académico.
Conclusión
En la presente comunicación presentamos un sistema de gestión de la información basado en herramientas de libre acceso disponibles para el personal sanitario. Un sistema que puede ser desplegado

de forma autónoma y completamente personalizada que nos permitirá explotar los datos derivados de la actividad de los ensayos clínicos en tiempo real, diseñando procesos de mejora en base a KPIs objetivos, obteniendo por tanto un mayor beneficio clínico de la actividad investigadora.

Transcriptome Translator: A free open-source Shiny App and R Function

Autor: Sonia María Rodríguez Huerta [1]

Co-autor: José Manuel Alvarez Díaz [1]

[1] *Universidad de Oviedo*

Within transcriptomic studies of Organism and Systems Biology, accurate data processing and translation are crucial for subsequent analyses. This work presents the development and release of a Shiny web application and an R function designed for transcriptome data translation. Additionally, an RMarkdown showcasing an example of the process conducted in the app and function is provided for enhanced transparency.

The Shiny application offers an intuitive interface for users to upload their transcriptome and download the translated results. Conversely, the R function provides a streamlined approach for automated translation processes.

Through seamless interaction with the Shiny application and/or the R function, researchers can effectively convert transcriptome data into translated formats suitable for various downstream analyses, thereby enhancing the accessibility and usability of transcriptomic data. This system represents a valuable resource for researchers seeking to harness the power of R programming and web-based applications in their transcriptome analyses.

Seguridad para las brigadas terrestres de extinción de incendios: cálculo con R de la ruta más rápida a un lugar seguro

Autor: Manuel Antonio Novo Pérez [1]

Co-autores: Marta Rodríguez Barreiro [1]; María José Ginzo Villamayor [1]

[1] *CITMAga*

Las brigadas terrestres de extinción de incendios forestales son un elemento muy importante en el control del incendio desde tierra. Durante su labor, se enfrentan al peligro de quedar cercados por el incendio. El protocolo OACEL (Observación, Atención, Comunicación, ruta de Escape y Lugar seguro) del Comité de Lucha Contra Incendios Forestales, establece que a lo largo de las labores de extinción de un incendio forestal se deben establecer rutas de escape por las que abandonar de forma segura el lugar de trabajo de las brigadas terrestres hacia una zona segura (en la que no hay peligro de ser alcanzado por el fuego ni un calor radiante excesivo). Debido a la naturaleza dinámica del incendio, estas rutas pueden cambiar y deben ser reevaluadas periódicamente, estableciendo nuevas rutas cuando sea necesario.
Se presenta una herramienta desarrollada con R que permite calcular una ruta de escape a pie para la evacuación de los medios de extinción terrestres desde su posición actual hasta una zona segura (que debe ser introducida por el usuario). El algoritmo desarrollado proporciona la ruta más rápida entre los dos puntos. Para el cálculo de esta ruta, se tiene en cuenta la vegetación, la pendiente del terreno, obstáculos que no pueden atravesar⋯Incluso se ofrece la posibilidad de que el usuario introduzca obstáculos de forma manual, que no podrán ser atravesados por la ruta. La ruta debe ir siempre en la dirección contraria al avance del incendio y mantenerse siempre a una distancia del mismo. También se presenta una aplicación de Shiny que permite ejecutar el algoritmo y visualizar la ruta de salida proporcionada.
Para el desarrollo del algoritmo, se hace uso de paquetes como httr para conectar con la API del

IDEE (Infraestructura de Datos Espaciales de España) para la obtención de la información de las aguas estancadas, o terra y sf para el tratamiento de datos espaciales.
Este algoritmo se ha desarrollado en el marco del proyecto CUI de la Agencia Gallega de Innovación (GAIN) de la Xunta de Galicia y la empresa Avincis Aviation Spain SA.

MoMo: un sistema automatizado de vigilancia de la mortalidad diaria integrando R

Autor: Inmaculada León Gómez [1]

Co-autor: Diana Gomez-Barroso [1]

[1] *Centro Nacional de Epidemiología (Instituto de Salud Carlos III)*

El Sistema de monitorización de la Mortalidad diaria por todas las causas (MoMo) forma parte del "Plan Nacional de actuaciones preventivas de los efectos de las temperaturas extremas sobre la salud" creado por el Ministerio de Sanidad. MoMo basa su información en registros de defunciones procedentes del Instituto de Estadística (INE) y de del Ministerio de Justicia (MJU) y de la Agencia Estatal de Meteorología (AEMET). El objetivo de este estudio es, describir los procesos automáticos de R que incluyen: ejecución de modelos estadísticos y publicación de resultados en los paneles Web de vigilancia de la mortalidad diaria. Estos procesos están integrados en una rutina Docker más amplia que además importa datos automáticamente y los almacena en bases de datos.
Métodos
Fuentes de datos: defunciones diarias por todas las causas del Instituto Nacional de Estadística (INE) y de los registros civiles automatizados del MJU, temperaturas de la AEMET y población del INE. Diariamente se ejecutan procesos automatizados en R para: Ejecutar el modelo estimativo y predictivo (Índice Kairós) de MoMo, que consiste en un modelo GAM paramétrico basado en regresión de Poisson multinivel por provincia, con tendencia y estacionalidad anual mediante splines, temperatura (dos variables sintéticas), y población como ofset. El modelo estimativo no incluye el año en curso ni el 2020 y elimina outliers, el modelo predictivo (Índice Kairós) incluye hasta el día en curso y no elimina outliers. Y Publicar los resultados de los modelos del sistema MoMo en la Web del Instituto de Salud Carlos III: https://momo.isciii.es/panel_momo/. Todos los procesos están integrados en tres entornos: desarrollo, preproducción y producción que permiten realizar pruebas y minimizar errores. La tecnología utilizada en todos estos procesos además de R (base, modelado, flexdashboard y shiny, R markdown) es: ftp, API, Python, procesos bash, Docker, Linux, html, css, js básico.
Resultados
MoMo proporciona estimaciones diarias de excesos de mortalidad por todas las causas y atribuibles a exceso o defecto de temperatura por sexo, grupos de edad, a nivel nacional, CCAA y provincia. Además el índice Kairós proporciona, para los mismos grupos, niveles de riesgo de mortalidad en el día en curso y predicciones en los cinco posteriores. En el pasado invierno, durante las semanas de máxima circulación de virus respiratorios (52/2023 a 04/2024) se estimaron 6173 exceso de defunciones por todas las causas y 547 atribuibles a defecto de temperatura.
Conclusiones
MoMo proporciona de forma automática y diaria estimaciones de excesos de mortalidad por todas las causas y atribuibles a variaciones extremas de temperatura, que permiten estimar de forma oportuna el impacto de situaciones de interés en Salud Pública en el contexto actual de cambio climático.

Tratamiento de datos composicionales incompletos en R con el paquete zCompositions

Autor: Javier Palarea-Albaladejo [1]

Co-autor: Josep Antoni Martin Fernandez [1]

[1] *Universidad de Girona*

Los datos composicionales se refieren a datos multivariantes representando partes de un total; típicamente expresados en unidades relativas como porcentajes, partes por millón, minutos/día, mg/L,

o similares. Su análisis requiere tener en cuenta esta naturaleza relativa, lo que se consigue de forma efectiva y bien fundamentada centrándose en log-cocientes entre las partes. En este contexto, los problemas de datos incompletos se corresponden comúnmente con la presencia de ceros que impiden la aplicación de la metodología log-cociente, pero también con valores faltantes en general. Los ceros a menudo coinciden con valores censurados derivados de la sensibilidad limitada de los instrumentos de medida (caso de datos continuos) o resultan de limitaciones del muestreo (caso de datos de conteo). El paquete zCompositions en R proporciona un entorno integrado para el estudio de patrones e imputación de ceros en distintos contextos en el marco del análisis log-cociente; y recientemente se ha extendido para abordar tanto el caso de datos faltantes como el caso de conjuntos de datos que incluyen ambos, ceros y datos faltantes simultáneamente. En este trabajo presentamos las características principales del paquete y algunos ejemplos ilustrativos de su uso en la práctica.

refseqR : operaciones computacionales comunes con registros de la colección RefSeq (NCBI)

Autor: Jose V. Die [1]

[1] *University of Cordoba*

La colección de secuencias de referencia del Centro Nacional de Información Biotecnológica (RefSeq, NCBI) mantiene un conjunto de secuencias completo, no redundante y bien anotado, que incluye genomas, transcritos y proteínas. En el momento de escribir estas líneas, el proyecto RefSeq contiene más de 60 millones de transcritos y 320 millones de secuencias de proteínas. En este trabajo describimos refseqR que proporciona un marco práctico para manejar secuencias biológicas alojadas en la colección RefSeq. refseqR simula el flujo de información genética dentro de un sistema biológico, permitiendo procesos direccionales desde loci genéticos recogidos como registros GenBank, a transcritos y de ahí a secuencias de proteínas curadas a partir de la base de datos RefSeq, así como otras combinaciones entre secuencias de estas moléculas. refseqR permite la interoperabilidad y la integración con varios objetos de Bioconductor proporcionando una conexión directa con otros proyectos. El paquete refseqR está implementado en R y se publica bajo la licencia MIT de código abierto.

Maximizando el Potencial del Análisis de Comunidades con R: Herramientas Integradas para la Evaluación de Artrópodos del Suelo

Autor: María del Carmen Fernández Bravo [1]

[1] *Universidad de Córdoba*

Los artrópodos edáficos representan una de las comunidades de organismos más diversos y complejos en los suelos de todo el mundo, los cuales contribuyen en la descomposición de la materia orgánica, el reciclaje de nutrientes, la mineralización del nitrógeno y el fósforo, la aireación y la diseminación de microorganismos, entre otras funciones. Muchos de estos artrópodos, concretamente la mesofauna, son especialmente sensibles a las perturbaciones naturales y antropogénicas de los ecosistemas, hecho que enfatiza su potencial como agentes para monitorizar la salud de los suelos.
La comunidad de usuarios y desarrolladores de R ha generado una amplia gama de paquetes y herramientas especializadas como iNEXT, Vegan o Indicspecies, entre otros, diseñados específicamente para abordar desafíos comunes en el análisis de comunidades. Estos paquetes ofrecen funciones específicas que simplifican tareas complejas, desde la estimación de la riqueza de especies, hasta la evaluación de la similitud comunitaria, la detección de indicadores de calidad de los ecosistemas o los complejos análisis multivariados y su modelización. Además, otra ventaja clave es la capacidad de generar visualizaciones claras y efectivas, que permite, por ejemplo, representar de manera intuitiva la superposición de especies entre diferentes condiciones ambientales.
El uso de R aplicado al análisis de comunidades de artrópodos de suelo proporciona una plataforma poderosa y completa para comprender la ecología de los organismos tan importantes. Su capacidad

para integrar análisis estadísticos avanzados con visualizaciones efectivas hace que sea una herramienta invaluable para investigadores y gestores del medio ambiente en su búsqueda por comprender y conservar la biodiversidad del suelo.

spTReg: Modelos de regresión bayesianos espacio-temporales, el caso de la regresión cuantil

Autor: Jorge Castillo-Mateo [1]

[1] *University of Zaragoza*

En campos tan diversos como las ciencias geológicas y medioambientales, la ecología, o la epidemiología es frecuente encontrar datos con referencia espacial recopilados a lo largo del tiempo en un conjunto fijo de localizaciones con coordenadas en una región de interés. El estudio de estos datos georreferenciados puede beneficiarse del marco de modelos jerárquicos bayesianos. Desafortunadamente, el ajuste de estos modelos pasa por resolver complejas integrales múltiples, y por tanto resulta inviable sin el uso de métodos computacionalmente intensivos difíciles de programar. En este trabajo se presenta el paquete de R **spTReg** disponible en GitHub (https://github.com/JorgeCastilloMateo/spTReg) y próximamente disponible en CRAN. Por el momento, el paquete proporciona un marco de modelos de regresión espacio-temporales para la media con errores gaussianos, para los cuantiles con errores asimétricos de Laplace, y para una respuesta binaria con link probit y logit. Los modelos operan sobre un espacio continuo y pueden adoptar hasta dos escalas temporales discretas, además pueden incorporar procesos gaussianos espaciales a distintos niveles para modelizar la dependencia espacial y autorregresión para modelizar la dependencia temporal. Los algoritmos de Markov chain Monte Carlo (MCMC) ofrecen una inferencia completa basada en el modelo y son los empleados para el ajuste de los modelos. La fuerte carga computacional de los algoritmos hace que sea muy beneficioso programarlos en un lenguaje de alto rendimiento como C++ cuyo código se conecta con **spTReg** mediante el paquete de R **Rcpp**. Además, también se hace uso del paquete de R **RcppArmadillo** que permite la conexión de R con el software de altas prestaciones de algebra lineal **Armadillo**. Finalmente, **spTReg** se ilustra mediante el estudio del efecto del cambio climático en los extremos de series de temperatura máxima diaria de la Península Ibérica por medio de una regresión cuantil espacio-temporal.

Interpretación de modelos de machine learning: estimando la importancia de las variables predictoras y los efectos marginales sobre la variable objetivo

Autor: Julio E Sandubete [1]

Co-autores: Marcial Fernández Amorós [1]; Silvia Gómez Hidalgo [1]; Lorenzo Escot [1]

[1] *Universidad Complutense de Madrid*

Al igual que durante el siglo pasado la inferencia estadística tuvo que desarrollarse para estimar modelos con los que conseguir extraer la máxima información de una muy corta cantidad de datos, la nueva realidad está caracterizada por disponibilidad de grandes cantidades de datos, y de diferente naturaleza que han dado origen a nuevos modelos y técnicas de análisis (machine learning, Deep learning, inteligencia artificial···). Estos nuevos modelos que podríamos calificar como de básicamente algorítmicos, y cuyo objetivo principal es la predicción, han venido a sumarse y a complementar, a los modelos de regresión tradicional, básicamente inferenciales y confirmatorios que utilizamos en las ciencias empíricas para entender el funcionamiento de los diferentes fenómenos de la realidad, el contraste de hipótesis y la cuantificación de la asociación entre variables dependientes e independientes.

Nos encontramos así ante un escenario de encuentro entre los dos enfoques disponibles para el análisis de datos, el puramente predictivo y algorítmico, por un lado, y el más explicativo e inferencial

por otro. Enfoques que representan las dos culturas de la modelización estadística a la que hacía referencia Leo Breiman (2001). Dos culturas que lejos de distanciarse y enfrentarse están siendo objeto de un acercamiento e integración. En efecto, esos modelos algorítmicos, principalmente predictivos, han sido denominados como de "caja negra", porque en ellos es complicado establecer de manera clara cuál es el efecto que tienen las diferentes variables predictoras o independientes sobre la variable objetivo o dependiente. Parte de los avances recientes en la Ciencia de Datos se encuentra precisamente en iluminar o dotar de luz a esos modelos de caja negra, para permitirnos avanzar en el mejor conocimiento del funcionamiento de la realidad.

En este trabajo queremos presentar de una manera práctica y aplicada algunas de las técnicas existentes para estimar el impacto y la importancia de las variables predictoras en los modelos de machine learning. Se presentarán librerías como DALEX o iml que tratan precisamente de dotar de interpretabilidad a estos modelos "de caja negra". Estas técnicas se compararán con otros desarrollos que mediante el uso de las expresiones algebraicas de las derivadas parciales en algunos de estos modelos posibilitan estimar el impacto marginal que cada una de las variables explicativas sobre la variable predictora.

CRAS: Aplicación Shiny para Análisis y Simulación de Riesgos de Ciberseguridad

Autor: Emilio L. Cano[1]

Co-autores: Carmen Lancho [1]; Víctor Aceña [1]; Marina Cuesta [1]; Rubén R. Fernández [1]; Isaac de Diego [1]

[1] *Universidad Rey Juan Carlos*

El **análisis y la gestión de riesgos** conllevan métodos tanto cualitativos como cuantitativos. Los métodos cuantitativos utilizados en el análisis de riesgos se basan en gran medida en sólidos métodos estadísticos y en la simulación de Montecarlo. La realización de análisis de riesgos cuantitativos y su comunicación se ha vuelto crucial en la Ciberseguridad, que afecta a todos los sectores, incluido el financiero. De hecho, los riesgos de Ciberseguridad se incluyen dentro de los "riesgos operativos" en el sector financiero.

En cuanto a la seguridad de la información, la metodología FAIR desarrollada por el Open Group se ha convertido en un estándar para el análisis de riesgos de ciberseguridad. FAIR utiliza distribuciones triangulares y PERT (Program Evaluation and Review Techniques) para simular y evaluar el impacto (pérdida) basándose en las aportaciones de los expertos. Sin embargo, pueden utilizarse otras distribuciones y métodos para analizar los riesgos, por ejemplo, distribuciones lognormales basadas en intervalos de probabilidad, entre otras.

En este trabajo, presentamos una aplicación shiny para la simulación de pérdidas en ciberseguridad, que permite al usuario elegir entre utilizar la metodología FAIR, o modificaciones de la misma, como diferentes distribuciones de probabilidad, o modificaciones sobre la ontología FAIR. El análisis estadístico de los resultados de la simulación se muestra mediante tablas y gráficos interactivos. Desde la propia aplicación, se puede generar automáticamente un informe y descargarlo para su uso posterior. El informe se genera en html, docx o pdf mediante una plantilla de documento quarto. La app también es útil para enseñar Análisis de Riesgos en cursos de grado y posgrado sobre ciberseguridad.

El trabajo futuro incluye añadir más distribuciones de probabilidad, como distribuciones de valores extremos, y publicar la aplicación como paquete en CRAN.

Fundamentos de ciencia de datos con R

Autores: Gema Fernández-Avilés Calderón [1]; José-María Montero [1]

[1] *University of Castilla-La Mancha*

El siglo XXI está siendo testigo de cambios vertiginosos en el contexto social y tecnológico, entre otros. Los tiempos han cambiado, la sociedad se ha globalizado y "exige" respuestas inmediatas a problemas muy complejos. Vivimos en el mundo de la información, de los datos, o mejor, de las bases de datos masivas, y los ciudadanos y, sobre todo, las empresas y los gobiernos, dirigen su mirada hacia el mundo científico para que les ayude a "oír las historias" que cuentan esos datos acerca de la realidad de la que han sido extraídos. Y dado su enorme volumen y sofisticación (en el nuevo renacimiento las imágenes y los textos, por ejemplo, también son datos), exigen algoritmos de nueva generación en el campo del machine learning, o incluso del deep learning, para "oír las historias" que cuentan. No parecen mirar al "antiguo" investigador científico, sino al "nuevo" científico de datos.

Es por ello imprescindible disponer de un manual con las características del que aquí se presenta: rigor, completitud, amplitud temática y variedad de perspectivas (está elaborado por más de cuarenta autores), y todo ello implementado de principio a fin con el software estadístico R.

EasyAmpR: genotipado de amplicones multilocus de secuenciación masiva en R

Autor: Miguel Camacho Sánchez [1]

Co-autor: Jennifer Leonard [1]

[1] *Estación Biológica de Doñana - CSIC*

La secuenciación masiva de librería de amplicones permite generar en paralelo datos para decenas o miles de loci y gran cantidad de muestras en un mismo experimento. Los sesgos, calidad de secuenciación y la gran cantidad de datos imponen algunos retos en el genotipado de muestras a partir de este tipo de datos genéticos. EasyAmpR es el primer paquete de R diseñado para automatizar el genotipado muestras de individuos diploides a partir de secuencias de librerías de amplicones de secuenciación masiva, tales como Illumina MiSeq o NovaSeq. Propone un flujo de trabajo que comienza con el demultiplexado de los loci, determinación de variantes con *dada2* y genotipado a partir de las variantes. Además, incluye varias funciones útiles para conversión entre formatos exportar resultados en forma de tablas y FASTA. Probamos exitosamente EasyAmpR en librerías multiplex de 30 loci para un estudio de genética de poblaciones de una especie de roedor del Sudeste Asiático, comparando las secuencias generadas por EasyAmpR con datos publicados para la misma especie.

CrossCarry segunda versión: Análisis de datos de un diseño cruzado mediante GEE

Autor: Nelson A. Cruz [1]

Co-autor: O. O. Melo [2]

[1] *Universitat de les Illes Balears*

[2] *Universidad Nacional de Colombia*

Los diseños cruzados experimentales se utilizan ampliamente en medicina, agricultura y otras áreas de las ciencias biológicas. Debido a las características del diseño cruzado, cada unidad experimental tiene observaciones longitudinales y presencia de efectos de arrastre sobre la variable respuesta. No

había ningún paquete en R que modelara claramente los datos de diseño cruzado. El paquete CrossCarry, cuya primera versión permitió analizar cualquier diseño cruzado siempre que la variable de respuesta observada pertenezca a la familia exponencial, independientemente de si existe o no un período de lavado. También permitió modelar mediciones repetidas dentro de cada período y amplió las estructuras de correlación utilizadas en ecuaciones de estimación generalizadas. Sin embargo, no se habían explorado las características para imputar los datos faltantes dentro de un diseño cruzado, pero ahora se ha propuesto y actualizado en el paquete una metodología para abordar estos problemas; Este es un avance significativo para el paquete y, más importante aún, para los diseños cruzados.

Además, ahora permite modelar líneas de base dinámicas y permite simulaciones de tamaño de muestra y pruebas de configuraciones de diseño cruzado antes de la aplicación del diseño en pacientes. Proporciona tablas de potencia esperada y posibles especificaciones erróneas de diseños basados en estudios farmacológicos. Estos últimos son muy importantes para el problema de los complejos efectos de arrastre, donde hemos logrado avances significativos en el último año y, por eso, esta nueva versión del paquete es más sólida y práctica.

Cruz, N. A., Melo, O. O., & Martinez, C. A. (2023). CrossCarry: An R package for the analysis of data from a crossover design with GEE. arXiv preprint arXiv:2304.02440.

Tang, Y. (2021). Power and sample size for GEE analysis of incomplete paired outcomes in 2× 2 crossover trials. Pharmaceutical Statistics, 20(4), 820-839.

Identificación de OTUs diferenciales en microbiomas: Extensiones de la función explore_logratios para clasificación binaria y multinomial

Autor: Irene García Mosquera [1]

Co-autores: Ricardo Alberich Martí [1]; Nelson A. Cruz Gutierrez [1]; Raquel Fernández Peralta [1]; Arnau Mir Torres [1]; Francesc Rosselló Llompart [1]

[1] *Universitat de les Illes Balears*

Presentamos dos extensiones de la función explore_logratios de la librería coda4microbiome (Calle et al., 2023) de R. Esta librería ha sido diseñada para identificar unidades taxonómicas operativas (OTUs) diferenciales de un microbioma mediante el ajuste de una regresión logística en todos los pares de log-ratios. Los coeficientes se estiman utilizando penalizaciones de los métodos de regresión ridge y lasso, conocidos colectivamente como elastic-net.
La primera extensión consiste en encontrar un conjunto de OTUs con log-ratio relevantes, tales que la media de las correlaciones log-ratio entre ellos sea máxima. Se propone una visualización de las asociaciones encontradas que permite detectar si hay máximos locales de interés. En la práctica, el conjunto de OTUs que da asociación máxima suele separar muy bien las dos categorías de la variable binaria de interés. Probamos esta extensión con varios conjuntos de datos de referencia, demostrando que proporciona un conjunto óptimo de OTUs para clasificar eficazmente las dos clases de la variable.
La segunda extensión permite utilizar explore_logratios en problemas de clasificación discreta no binaria. Esta metodología se basa en el área bajo la curva (AUC) de predicción multinomial, como se describe en (Hand y Till, 2001) y permite emplear la primera extensión para la identificación de log-ratios relevantes en contextos con más de dos categorías.

Using Recurrence Analysis to search for patterns of dynamic behavior in economic time series

Autores: Elena Olmedo [1]; Lorenzo Escot [2]; Iñaki Aliende [2]

[1] *Universidad de Sevilla*

[2] *Universidad Complutense de Madrid*

The Recurrence Analysis is a statistical technique, initially proposed by Eckmann et al. (1987) and Zbilut and Webber (1992), used to analyze dynamic systems and time series. This technique is based on the identification and study of repetitive patterns (recurrences) within a time series, which can provide information about the underlying structure of the system or the process that generated the data. In this paper we will use the crqa (Cross Recurrence Quantification Analysis) library, which allows us to perform both recurrence plots and recurrence quantification analysis. Some essential concepts to understand this technique such as phase space reconstruction, embedding dimension or reconstruction lag will be introduced beforehand. We will show some examples of application of recurrence analysis with time series whose generating process is well known in order to later apply it to the analysis of different observed economic time series in order to detect patterns of dynamic behavior that allow us to complete (not replace) the traditional analysis of correlation or time dependence of these time series.

En busca de la eficiencia: el paquete 'labeleR' como método automático y flexible para generar etiquetas y certificados

Autor: Ignacio Ramos-Gutiérrez [1]

Co-autores: Francisco Rodríguez Sánchez [2]; Jimena Mateo-Martín [1]; Julia G. de Aledo [1]

[1] *Universidad Autónoma de Madrid*

[2] *Universidad de Sevilla*

El paquete 'labeleR' ha sido desarrollado para facilitar la generación de etiquetas de especímenes de herbario y otras colecciones científicas, certificados de asistencia y participación, acreditaciones identificativas. Esta herramienta permite la renderización automática de etiquetas y documentos a partir de extensas bases de datos, optimizando procesos típicamente largos y repetitivos. Este paquete se presenta como una opción personalizable y flexible frente a otras alternativas de uso frecuente, como la función mailing list combinando Microsoft Excel y Word, las funciones de Microsoft Access o programas informáticos alternativos que suelen ser caros y requieren conocimientos previos.

En su lugar, el paquete labeleR ha sido desarrollado en código abierto, lo que permite descargarlo gratuitamente desde el repositorio CRAN o desde GitHub (donde se puede acceder a un tutorial detallado, ver https://ecologyr.github.io/labeleR/). Los usuarios pueden fácilmente elegir su tabla de datos de origen, introducir códigos QR y logotipos, modificar el contenido según sus necesidades específicas e incluso crear sus propias plantillas personalizadas. El paquete tiene seis funciones principales, tres de ellas centradas en la gestión de colecciones científicas (etiquetas de herbario, etiquetas de colección y una versión más pequeña llamada "tinylabels"), y otras tres funciones en la organización de eventos científicos (acreditaciones para eventos, certificados de asistencia y certificados de participación). El paquete labeleR convierte un proceso tedioso en una actividad rápida para obtener etiquetas y/o certificados en formato PDF para su fácil impresión o distribución. Creemos que esta versátil herramienta contribuye a la eficiencia científica al simplificar procesos de etiquetado y organización complejos y que exigen mucho tiempo.

max_clique: the project and development of a new package for robust clustering

Autor: Stefano Benati [1]

Co-autor: Alessandro Avellone [1]

[1] *Università di Trento*

The clique partitioning problem is a classic integer linear programming model: Problem data are a set of units U, characterized by a relation of similarity/dissimilarity d(i,j) between every pair of units i,j in U. Set U must be partitioned in such way that the sum of the similarities of the pairs (i,j) that are in the same group is as high as possible. When we compare the clique partitioning problem to other computational tools for clustering, like the k-means for example, we can recognize some advantages.

1. Clique partitioning can be defined for both qualitative and quantitative data, as only an appropriate similarity measure must be defined. Conversely, the k-means, requiring averages, can work only with quantitative data.
2. The number of clusters is not defined by the user, as it is a problem output. The clique model itself selects the optimal group of clusters. Conversely, the k-means is usually run with varying values of k, then the correct k is selected by some rule-of-thumb.
3. Consequence of point 2, outliers are detected by singletons of the partition. Conversely, outliers of the k-means model can be detected only by some preliminary analysis.
4. The clique partitioning model is flexible enough to be used for community detection too. That is, a clustering model in which units are connected by links. In this framework, the model is called modularity maximization.
5. Clique partitioning and k-means are both NP-complete. However, clique is integer programming, while k-means is non-convex programming. It is easier to find the optimal solution of an ILP, rather than to a non-convex problem.

Even though there are many reasons to prefer a clique model to a k-means, still the latter is more known and popular. One of the reasons is the lack of a reliable procedure and package for clique partitioning in the most important platforms: R, Julia or Python. In our communication, after revising the combinatorial and statistical properties of the clique partitioning model, we will illustrate our computational experience with it and describe the structure of new R package, under development for our research projects. Our package is composed of heuristic and exact procedures, some new and some already known in the literature. Nevertheless, an original package feature is the use of the set partitioning formulation of the clustering model. This feature is a unifying framework that can be used to decompose various models into a master and a pricing problem. While the master problem is the same for many clustering models, pricing is peculiar. It implies that users can test new clustering model and interact with the package defining only their own pricing problem.

Development of our own R packages for the statistical processes of the Generic Statistical Business Process Model

Autor: Elisa Jorge [1]

Co-autor: Alberto González Yanes [1]

[1] *Instituto Canario de Estadística (ISTAC)*

La estadística pública se enfrenta a una demanda cada vez mayor de productos estadísticos que sean oportunos, frecuentes y precisos. La mejora continua de los procesos estadísticos, incluyendo la estandarización y automatización, se hace cada vez más evidente para poder mantenerse al día con dicha demanda y al mismo tiempo garantizar que se cumplan los estándares de calidad. Por ello, surge la necesidad de implementar herramientas TI que puedan usarse para una amplia gama de estos procesos, que sean rentables y cuenten con una amplia aceptación y soporte.

Bajo esta premisa, el Instituto Canario de Estadística (ISTAC) ha estado trabajando en la elaboración de su propia librería de uso interno para la estandarización y automatización de sus procesos enmarcados dentro del Modelo Estadístico Genérico de Procesos de Negocio (GSBPM). Además de fomentar el uso de R internamente, el ISTAC también contribuye activamente a la comunidad internacional de código abierto mediante el desarrollo de librerías de interés para las estadísticas oficiales.

Esta presentación tiene como objetivo resaltar la importancia del uso de un código común, sostenible y reutilizable, así cómo exponer su uso en algunas de las diferentes fases y subprocesos del GSBPM en las actividades del ISTAC y su implementación.

Analysis of the RFSI interpolation method for the precipitation variable in Spain

Autor: Lorena Galiano Sánchez [1]

Co-autores: Carlos Prado López [1]; Darío Redolat Negro [1]; Dominic Royé [1]; Robert Monjo [1]

[1] *Fundación para la Investigación del Clima*

Spatial interpolation methods offer the ability to obtain the regional spatial distribution of a variable from local measurements, as long as there is a sufficient density of observatories to do so. It should be noted that the interpolation process takes into account the terrain factors that influence the variable within the study region. In this way, representative values are obtained for each point in a given region, allowing for a complete graphical representation.

The precipitation variable has a large spatial variability even between relatively close points due to local geographic effects and, in turn, this variability can vary depending on the time of year in a seasonal precipitation regime. Therefore, it is necessary to apply interpolation methods that assess the relationship of predictors to the prediction in each specific case. This study presents the analysis of the Random Forest Spatial Interpolation (RFSI) method applied to the precipitation variable in a study area covering the Iberian Peninsula and the Balearic Islands. For this purpose, the "meteo" library developed for R is used. The results obtained by using different combinations of predictor variables are evaluated. In addition, the results are compared with other benchmark interpolation techniques, such as the generalised linear fit model.

Coexpresión y selección de variables con submuestreo aleatorio

Autor: Amalia Cristofaro [1]

Co-autores: Andrea Sánchez-Moreno [1]; David Barrios [1]; Carlos Prieto [1]

[1] *Universidad de Salamanca.*

Las técnicas de selección de variables son esenciales en biomedicina para indentificar variables clave en bases de datos clínicas y para aplicar técnicas con poder clasificatorio y predictivo. En este trabajo se presenta una nueva aproximación basada en submuestreo aleatorio que se puede aplicar en técnicas de selección de variables para conseguir métodos estadísticos robustos. Esta técnica es apropiada en proyectos con un gran número de muestras y tiene una aplicación directa en proyectos de expresión de célula única. Como ejemplo mostraremos los resultados obtenidos en un problema de predicción de la edad mediante el análisis de datos de expresión de ARN.
Para su aplicación hemos desarrollado en paquete GeneCoexp, que además de realizar una selección de características con técnicas de coexpresión, permite identificar relaciones de coexpresión robusta entre genes y diferenciarlas de las producidas por una posible estratificación de las muestras. Para su ejecución es necesario realizar un gran número de submuestreos, por lo que ha sido necesario

reimplementar las funciones de coexpresión y usar paralelización para conseguir resultados en tiempo real. Como resultado el paquete permite generar una red de coexpresión robusta que se puede explorar interactivamente con el paquete rD3plot.

Financiación
A.C. ha sido financiada por el "Programa investigo" en el marco del Plan de Recuperación, Transformación y Resiliencia (SEPE, JCyL, Fondos NextGenerationEU financiados por la Union Europea). A.S. recibió inanciación del Programa Operativo de Empleo Juvenil, Fondo social Europeo (FSE), Junta de Castilla y Leon (JCyL). D.B. ha sido financiado por el programa PTA (PTA2022-022270-I) del Ministerio de Ciencia, Innovación y Universidades.

El efecto de factores socioeconómicos sobre la conectividad del transporte público

Autor: Javier Hernán Matas Monroy [1]

Co-autores: Juan María Hernández Guerra [1]; Rafael Ricardo Suárez Vega [1]

[1] *Universidad de Las Palmas de Gran Canaria*

Este trabajo de investigación tiene como objetivo analizar el grado de conectividad a través del transporte público de los distintos núcleos poblacionales de la isla de Gran Canaria con diversos polos de atracción social, como por ejemplo las universidades, aeropuertos, parques de atracciones, entre otros. La principal finalidad es estudiar la relación entre las características sociodemográficas de los núcleos poblacionales y su conectividad con los polos de atracción.
En este contexto, el software R ha permitido manipular la base de datos proporcionada por la empresa operadora del transporte público en la isla y disponerla en formato de red (con pares de observaciones origen-destino). Además, a partir de las llamadas a APIs de Google usando diversas librerías disponibles, se pudieron obtener todos los tiempos de viaje y distancias entre cada una de las paradas incluidas en la red. Más adelante hemos hecho uso de R para la representación geográfica de todas las estaciones de autobuses incluidas en la base de datos, con el objetivo de identificar qué paradas estaban asociadas a cada conjunto poblacional. Por último, se estimaron modelos de regresión espacial para determinar qué factores tenían influencia sobre el grado de conectividad de cada localidad, así como el tipo de efecto que se producía.
En resumen, a partir del uso de R hemos manipulado una base de datos hasta convertirla en formato grafo, se ha representado la red obtenida, se ha podido incluir datos de tiempo de viaje y distancia entre dos puntos a partir de llamadas a APIs de Google, y se han estimado modelos espaciales, así como su análisis referido a la bondad del ajuste de los mismos.

Cuentas nacionales a gran escala con R: FIGARO multi-country input output tables

Autor: Kamal Antonio Romero Sookoo [1]

Co-autores: David Barta [2]; José Manuel Rueda Cantuche [3]; Ricardo Lobato [4]; Santacruz Banacloche Sánchez [3]

[1] *External consultant JRC European Commission*

[2] *External consultant Eurostat*

[3] *JRC European Commission*

[4] *Eurostat*

Las tablas FIGARO de Eurostat proporcionan una visión detallada de las transacciones económicas entre sectores y países. En esta presentación, se describirá el proceso de construcción de estas tablas utilizando el lenguaje de programación R. Utilizar R para la elaboración de las tablas FIGARO permite no solo una gestión eficiente de datos complejos, sino la implementación rápida de algoritmos de

balanceo y una evaluación analítica y gráfica de los resultados en forma de reportes automatizados. En la presentación abordaremos las distintas áreas del proceso como la recopilación y transformación de ingentes datos, el cálculo de tablas intersectoriales multi-país, y la validación y visualización. Se discutirán los desafíos encontrados durante el proceso y las soluciones implementadas, destacando cómo R facilita un flujo de trabajo flexible y reproducible.

Grafos de coincidencias y regresión

Autores: Modesto Escobar [1]; Cristina Calvo López [1]

Co-autores: Carlos Prieto [1]; David Barrios [1]

[1] *Universidad de Salamanca*

Los gráficos se han empleado no sólo para resolver problemas topográficos y representar estructuras sociales, sino también para mostrar la correlación entre variables según modelos casuales. De hecho, el análisis de trayectorias y los modelos de ecuaciones estructurales son bien conocidos por los científicos sociales, pero ambos se limitaron a variables cuantitativas en sus primeras etapas. Con el paquete de R netCoin se pueden obtener nuevas formas de mostrar las conexiones entre variables cualitativas de forma similar al análisis de correspondencias, pero utilizando otro conjunto de técnicas multivariantes, como la regresión lineal y logística, mezcladas con el análisis de redes.
El NCA (Network Coincidence Analysis) se puede emplear particularmente para el análisis exploratorio de datos de encuestas. Para ello, los nodos representan las distintas categorías de las variables seleccionadas, mientras que los enlaces simbolizan las relaciones entre las distintas variables. Uno de los usos específicos de esta técnica de análisis consiste en la caracterización mediante diversas variables sociodemográficas de diferentes perfiles de respuesta. Además de las medidas de correlación, el análisis propuesto puede estimar modelos log-lineales para estudiar las relaciones multivariantes, incluidas las interacciones. Incluso, para aumentar la potencia analítica de estas herramientas, disponen de características interactivas en línea, que incluyen tanto la selección de los elementos en función de su tamaño o atributos, como el filtro de los vínculos más centrales y fuertes.

La primera parte de la presentación trataría de la base estadística de estas representaciones y la segunda daría ejemplos del uso del paquete en encuestas comparativas internacionales como la Encuesta Social Europea.

Estimación del exceso de defunciones durante la pandemia de COVID-19 con R

Autor: Virgilio Gómez Rubio [1]

Co-autores: Garyfallos Konstantinoudis [2]; Michela Cameletti [3]; Monica Pirani [2]; Gianluca Baio [4]; Marta Blangiardo [2]

[1] *Universidad de Castilla - La Mancha*

[2] *MRC Centre for Environment and Health, Imperial College London*

[3] *Department of Economics, University of Bergamo*

[4] *Department of Statistical Sciences, University College London*

La pandemia de COVID-19 produjo un exceso de defunciones en varias regiones del mundo. En mi charla ilustraré qué necesitamos y cómo podemos estimar este exceso de defunciones con R. En particular, hablaré de todo el flujo de trabajo desde la carga de datos hasta la visualización de los resultados pasando por el análisis de los datos de población y mortalidad utilizando modelos espacio-temporales.

La presentación estará basada en estos artículo publicados en el R Journal y Nature Communications:

Konstantinoudis, et al., "A Workflow for Estimating and Visualising Excess Mortality During the COVID-19 Pandemic", The R Journal, 2023, https://doi.org/10.32614/RJ-2023-055 .

Konstantinoudis, G., Cameletti, M., Gómez-Rubio, V. et al. Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. Nat Commun 13, 482 (2022). https://doi.org/10.1038/s41467-022-28157-3

Una aplicación de datos abiertos en la Administración pública: El Panel de Indicadores de Turismo de Lanzarote

Autor: José Manuel Cazorla Artiles [1]

Co-autor: Christian González Martel [1]

[1] *Universidad de Las Palmas de Gran Canaria*

En este trabajo se muestra una aplicación de cómo utilizar la información estadística pública del Instituto Canario de Estadística (ISTAC) y el lenguaje de programación R para aumentar el conocimiento del sector turístico en la isla de Lanzarote.

El resultado es el Panel de Indicadores de Turismo de Lanzarote, una aplicación web desarrollada en R, empleando especialmente la librería shiny. Esta aplicación se actualiza diariamente haciendo uso de la API del ISTAC.

Adicionalmente, y con objeto de aumentar la difusión de los resultados, se ha desarrollado la automatización de informes que reflejan la coyuntura del sector turístico en la isla de Lanzarote.

Integrating R Software as a Teaching Resource for STEM Education: A Multidisciplinary Approach

Autor: Xavier Angerri [1]

Co-autores: Karina Gibert [1]; Sergi Ramirez-Mitjans [1]

[1] *KEMLG at Intelligent Data Science and Artificial Intelligence (IDEAI-UPC)*

In the context of higher education, teaching Science, Technology, Engineering and Mathematics (STEM) disciplines constantly faces the challenge of staying relevant and effective in a world characterized by rapid technological advancements and socioeconomic changes. In particular teaching advanced statistics and data science or machine learning have many challenges due to the complexity of the methods and the limitations for formal demonstrations in non-maths carrears. In response to this demand, this project proposes an innovative methodology that uses R software and some of their specific libraries as the main teaching resource in advanced statistical and data science training for STEM higer education.
R software 1, widely used in the scientific and business community for data analysis and visualization, offers a rich and versatile environment for exploring and applying key concepts in STEM degrees. By adopting R as a pedagogical tool as it is shown in 2, the aim is not only to teach students the technical skills necessary for data manipulation and statistical analysis but also to foster a deeper understanding of the underlying principles in areas such as data science, mathematical modeling, and programming.
The proposed methodology is based on a practical and experiential team work approach, where students learn by doing as discussed in 3.

The methodology is based on building wide groups of students to face long term project for an entire term where they will ve asked to apply the theoretical knowledge gained in their own real data, enabling them to find the difficulties of real data and assisting them on overcomming the challenges week by week. This contributes to
develop transferable skills and a more solid understanding of concepts. Every week the theoretical class introduces a concept from data collection, preprocessing, modelling or validation and the corresponding practical class consists on applying the new knowledge to their project. Lecturers provides a basic script in R, RMarkdown or Shiny depending on the week . The script contains some errors and they have to tune it to their particular data by correcting the mistakes as well . Sometimes the lecturer runs the script with some academic data while in parallel they try to run with their own data. Thanks to this pedagogic technique, the students learn to solve mistakes in the code, gain practical skills on real applications and connect what they do with the theoretical concepts. Additionally, the focus on collaboration and teamwork promotes peer learning and fosters a supportive and collective environment in the classroom. Lecturers answer questions and doubts during lab sessions and guides the group as far as they can go.
The projects are presented in public sessions to the entire class twice or 3 times along the course so that the evolution of the projects is progressive.
The common discussion is part of the training programm as well. Each dataset has different requirements at the level of preprocessing, modelling or interpretation and sharing the work to the entire class allows to learn from the experience of other groups as well. This is relevant as the knowledge corpus of these courses is huge and this is a way to cover more topics inside a term.
One of the distinctive aspects of this methodology is its ability to integrate multiple STEM disciplines into a single course or project. By using examples and applications that span diverse areas such as biology, engineering, economics, and social sciences, students have the opportunity to explore the interconnections between different fields and understand how they can apply their skills in a variety of contexts.
However, despite the many potential benefits that this methodology offers, it also presents unique challenges. The initial learning curve for students can be steep, especially for those with no prior experience in programming or data analysis. Additionally, the effective integration of R into the curriculum requires careful planning and adequate resources, including teacher training and access to suitable hardware and software.
As a conclusion, this project proposes an innovative methodology for teaching STEM disciplines that uses R software as the main learning engine. By providing students with a practical and multidisciplinary experience, it is hoped that this methodology will not only prepare students for real-world challenges but also inspire a lasting passion for learning and exploration in the field of STEM.

Palabras claves. Education, R, Statistics methods, Machine Learning

References
1 de Micheaux, P. L., Drouilhet, R., & Liquet, B. (2013). The R software. Fundamentals of Programming and Statistical Analysis, 978-1.
2 Koretsky, M., Keeler, J., Ivanovitch, J., & Cao, Y. (2018). The role of pedagogical tools in active learning: a case for sense-making. International journal of STEM education, 5(1), 1-20.
3 Thompson, P. (2010). Learning by doing. Handbook of the Economics of Innovation, 1, 429-476

Fuzzy Logic System for Determining Emotional and Mental States in R

Autor: Victoria López-López [1]

Co-autor: Roberto Morales-Arsenal [1]

[1] *CUNEF Universidad*

One in four people worldwide will experience a mental disorder at some point in their lives, primarily related to anxiety, sleep disorders, or depression. In this work a fuzzy algorithm has been developed in R to determine the emotional and mental state of patients. In particular, we focus on measuring manic symptoms. For this task, a total of 11 fuzzy variables are defined according to the Young

scale: hyperactivity, euphoria, sexual impulse, sleep, irritability, verbal expression, thought process disorder, formal thought disorders, aggressiveness, appearance, and awareness of illness. First, we choose the real variable (measurable by sensors) that is associated with each fuzzy variable. For example, the fuzzy variable 'hyperactivity' can be associated with the real variable number of steps measured by an accelerometer. The fuzzy variable 'sleep' can be associated with the real variables produced by sound and motion sensors. For each fuzzy set, category of the variable, there is an associated membership function for its elements, which indicates to what extent the element is part of that fuzzy set. The most typical forms of membership functions trapezoidal, linear and curved are used. The results obtained show that the use of fuzzy classification systems improves the understanding of the patient compared to traditional classification methods that use Boolean systems.

LobbyBot: Análisis y Clasificación Automatizada de las Estrategias de los Grupos de Interés en las Noticias de los Medios de Comunicación de España

Autor: Aritz Gorostiza [1]

Co-autores: Antonio Castillo [1]; Encarna Hidalgo [2]

[1] *Universidad de Málaga*

[2] *Universidad de Granada*

Esta investigación presenta la segunda fase de "LobbyBot", una herramienta desarrollada en R y presentada en el II Congreso de R y XIII Jornadas de Usuarios de R en Barcelona en 2023. En esta segunda fase, nos enfocamos en la construcción de un modelo de aprendizaje automático para clasificar noticias según la estrategia comunicativa empleada por los grupos de interés que aparecen en ellas. Para ello, se desarrollará un modelo de aprendizaje automático utilizando el paquete CARET en R. Este paquete (abreviatura de Classification And REgression Training) es un conjunto de funciones que intenta simplificar el proceso de creación de modelos predictivos. La base de datos está compuesta por 431,797 noticias recolectadas en el último año mediante "LobbyBot", de las cuales, en aproximadamente 5,768, se menciona a algún grupo de interés registrado en la CNMV. El corpus lingüístico obtenido se dividió en un conjunto de entrenamiento (75%), que será etiquetado manualmente por los investigadores para validar el modelo con los datos restantes. Los resultados ofrecerán una visión detallada sobre las dinámicas de las estrategias comunicativas de los grupos de interés en los medios de comunicación españoles. Además, demostrarán la capacidad de las técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático (ML) para analizar las estrategias de comunicación predominantes según la naturaleza de cada organización en los medios de comunicación de España.

Optimización, Solvers y R con ejemplos

Autor: Alberto Torrejón Valenzuela [1]

[1] *SevillaR*

Al optimizar, como su propia definición indica, buscamos la mejor manera de realizar una actividad ahorrando recursos y tiempo. La optimización es un campo de las matemáticas que engloba un gran número de problemas, desde la optimización de horarios, localización de recursos de forma óptima, búsqueda de rutas óptimas que minimicen el tiempo de viaje, etc. Además, la optimización trasciende a otros campos, por poner algunos ejemplos, es útil para la mejora de algoritmos de aprendizaje automático o se usa en problemas de redes complejas para encontrar estructuras u otros patrones de interés. En este taller veremos como plantear problemas de optimización con programación matemáticas y resolver estos modelos en R con ayuda de solvers, softwares que permiten calcular la solución de estos modelos.

Diseño de experimentos: error absoluto vs error relativo constante

Autor: Carlos de la Calle Arroyo [1]

Co-autores: Chiara Tommasi [2]; Licesio Jesús Rodríguez-Aragón [3]; Samantha Leorato [2]

[1] *Universidad de Oviedo*

[2] *Università degli studi di Milano*

[3] *Universidad de Castilla-La Mancha*

Este trabajo se centra en mejorar las metodologías de diseño mediante la introducción de puntos de apoyo suplementarios, con el objetivo explícito de garantizar un nivel mínimo de KL-eficiencia para una selección óptima entre varias especificaciones de varianza. La metodología se basa en la extensión de la metodología de diseño D-aumentados para el criterio de KL-optimalidad. Además, esta estrategia resulta beneficiosa para modificar diseños existentes, como los diseños D-óptimos, para abordar el desafío de la especificación precisa de la varianza del error. Estos dos enfoques se comparan con el 'gold standard' en diseños multiobjetivo, que son los criterios compuestos, los cuales en este problema requieren la implementación de técnicas algorítmicas más complejas, como las metaheurísticas. Las ventajas e inconvenientes de las dos metodologías se valoran a través de una aplicación de las mismas a un problema pertinente.

Buscando alternativas a los índices de confort térmico en exteriores

Autor: José Antonio Rodríguez Gallego[1]

[1] *Universidad de Sevilla*

El estudio del confort térmico es un campo complejo, que ha evolucionado significativamente, y que muestra comportamientos de alta variabilidad especialmente cuando consideramos este fenómeno en entornos exteriores. Se ha considerado tradicionalmente que la percepción del confort se ve influenciada por factores microclimáticos tales como la temperatura del aire, la humedad, el viento, la radiación solar; pero en los últimos años se ha comprobado en repetidas ocasiones que dichas variables no explica con suficiente precisión este fenómeno, siendo necesario considerar factores personales. Este problema se hace más evidente al considerar la subjetividad en la percepción, que introduce mucho ruido en los estudios realizados, complicando considerablemente su análisis.

Se han desarrollado diversas herramientas para evaluar el confort térmico exterior, encontrándose entre los índices más destacados Índice de Temperatura Fisiológica Equivalente (PET) y el Índice Universal de Clima Térmico (UTCI). El PET, desarrollado en los años 90, simula la respuesta del cuerpo humano a las condiciones ambientales; mientras que el UTCI, introducido en 2009, considera una amplia gama de parámetros meteorológicos y fisiológicos para una evaluación más comprensiva. Ambos índices, sin embargo, presentan limitaciones en precisión y en capturar la complejidad de la percepción subjetiva, ignorando variables personales o relativas a la morfología del entorno.

Como respuesta a estas limitaciones, se está desarrollando una alternativa basada en técnicas de *machine learning*, utilizando diversos modelos para mejorar la precisión de las predicciones de confort térmico. A pesar de los avances, los modelos tradicionales de *machine learning* aún enfrentan desafíos significativos debido a la complejidad y subjetividad del fenómeno, cuestión que resulta además especialmente difícil de abordar debido a la ausencia de datos de calidad de dimensiones apropiadas para el entrenamiento de estos modelos.

Ensamblaje de bases de datos con validación automática utilizando GitHub Actions

Autor: Francisco Rodríguez-Sánchez [1]

[1] *Universidad de Sevilla*

El ensamblaje de bases de datos a partir de fuentes diversas es una tarea común que a menudo se realiza manualmente. Este proceso de ensamblaje manual no solo es lento e ineficiente, sino que además propicia la acumulación de errores. En esta charla se demostrará un flujo de trabajo para el ensamblaje de bases de datos con control de calidad automático utilizando GitHub Actions (computación en la nube). Dicho ensamblaje puede combinarse con la actualización dinámica y automática de páginas webs, posts en redes sociales, etc. de manera eficiente y totalmente gratuita.

Algoritmos basados en Generación de Columnas y Branch-and-Price

Autor: Francisco Temprano García [1]

[1] *Universidad de Sevilla*

Dentro de los muchos algoritmos y métodos para resolver de forma exacta y heurística un problema de optimización combinatoria, aquellos basados en técnicas de descomposición han tomado una gran fuerza e importancia en los últimos años. Una de estas técnicas es la generación de columnas, la cual nos permite reducir en gran medida la dimensión y tamaño de un problema de programación matemática cuyo número de variables puede ser exponencial.

A pesar de que esta técnica aun no ha sido implementada en la gran mayoría de solvers comerciales, con cierta noción de programación matemática y la ayuda de estos solvers para resolver subproblemas más sencillos, es posible diseñar un algoritmo Branch-and-Price para resolver de forma exacta problemas de optimización con una alta complejidad (NP-Completos).

BayesianNetworks: a new tool for analysing ecological interactions

Autor: Elena Quintero Borrero [1]

Co-autor: Francisco Rodríguez-Sánchez [1]

[1] *Universidad de Sevilla*

Studying species interactions is crucial for understanding ecosystem dynamics. However, obtaining accurate estimates of biotic interactions requires extensive field sampling over long periods, where more effort leads to more comprehensive estimates. To date, most empirical ecological studies are based on observational data which does not effectively incorporate sampling completeness into species interaction estimates. Yet, strong variation in sampling effort among species can greatly impact network structure as well as many other network parameters. Variation encountered across network structure and parameters can be a result of methodological bias rather than biological processes, hindering correct ecological interpretation. This underscores the need to consider uncertainty when dealing with unreliable ecological data. In a recent study, Young and colleagues (2021) proposed a Bayesian statistical framework that allows obtaining more robust estimates of network structure and ecological metrics from noisy observational data. The authors reconstructed plant-pollinator networks from observational data based on interaction likelihood. Building on the work of Young et al. (2021), we introduce a new open-source R package called "BayesianNetworks" that greatly

facilitates data preparation, model fitting, and posterior model assessment for large numbers of interaction networks. This package incorporates the effect of varying sampling efforts on network reconstruction, as well as underlying preferences between interacting partners. The package will be made publicly available soon and is expected to facilitate the adoption of a more robust framework for the analysis of ecological networks. We illustrate the application of this R package using 46 plant-animal networks based on fruit-consumption visitation events with varying sampling efforts. For each network, we obtain Bayesian posterior estimates for all potential interactions and propagate these uncertainties into network descriptors. We demonstrate how this novel confidence-based approach allows obtaining more robust, complete, and insightful picture of the structure of ecological networks.

Young, J.-G., Valdovinos, F. S., & Newman, M. E. J. (2021). Reconstruction of plant–pollinator networks from observational data. Nature Communications, 12(1), 3911. https://doi.org/10.1038/s41467-021-24149-x

GBS - Una aplicación Shiny-webpara pedir cambio en las máquinas expendedoras de billetes

Autor: Tobias Kellner [1]

[1] *Transdev Germany*

El 'GBS' es una aplicación web 'Shiny' que permite pedir nuevo cambio y papel para más de 2000 máquinas expendedoras de billetes. La aplicación que desarrollé lleva un año en uso productivo en mi empresa. No sólo ha supuesto un ahorro de varias horas de trabajo al día, sino que también ha ahorrado a la empresa varios cientos de miles de euros.
La interactividad y la facilidad de uso son el centro del desarrollo con el lenguaje de programación R. Pero en la presentación también se tratarán los retos de los procesos 'backend'. En la primera parte se presentarán los procesos necesarios para extraer los datos y darles el formato adecuado. La segunda parte de la presentación está dedicada al desarrollo en R-Shiny, mediante el cual se superaron los problemas de programación con funciones especialmente desarrolladas.

Efecto de la poliploidía sobre las redes de coexpresión de genes

Autor: Alex Oliva Fernández [1]

Co-autores: Francisco J Balao Robles [1]; Inmaculada Barranco Chamorro [1]

[1] *Universidad de Sevilla*

La poliploidía es el fenómeno en el que un organismo presenta varias dotaciones completas de cromosomas y puede surgir por distintos procesos, como la duplicación del genoma completo (WGD). Esta duplicación puede afectar significativamente a la expresión génica y las interacciones entre genes, ya que en lugar de duplicar los productos génicos estos tienden a desbalancearse. Las redes de coexpresión génica (GCN) relacionan los genes según la relación entre sus patrones de expresión, lo cual puede dar información sobre la función de los genes y las vías de señalización que regulan la expresión. Se hipotetiza que tras eventos de WGD habrá módulos con genes que no dupliquen su expresión junto con su dosis y que esto altera significativamente la estructura de la red, especialmente si dichos genes tenían alta conectividad. En este estudio se hace uso de las herramientas de construcción y alineamiento de redes de correlación de la librería WGCNA para evaluar la conservación de la estructura de las GCN en eventos de duplicación.

Con este fin se extraen subredes por módulos de la GCN de organismos diploides de Dianthus Broteri y se compara su estructura con la de las subredes con los mismos genes en datos reales y simulados de organismos tetraploides en 5 grupos distintos. Los datos reales corresponden a organismos del

mismo experimento, que han sufrido WGD de forma natural o causada en el laboratorio. Se simulan los niveles de expresión tras eventos de duplicación sobre los organismos diploides del experimento en los que se desbalancea la expresión de ciertos genes en dos de los grupos restantes. Finalmente se realiza un análisis estadístico de la similaridad de la estructura de los módulos para evaluar las hipótesis biológicas.

Moving forward in developing an innovative application for monitoring heat-related mortality in Spain.

Autor: Dominic Royé [1]

Co-autores: Aurelio Tobías [2]; Carmen Iñiguez [3]

[1] *Fundación para la Investigación del Clima (FIC)*

[2] *Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC)*

[3] *Universidad de Valencia*

Exposure to heat poses a major threat to high-risk populations by substantially contributing to increased morbidity and mortality. Heat-related mortality has been a significant concern since the extreme summer of 2003, when Europe experienced a heatwave, leading to an excess of more than 70,000 deaths during the summer months. In the context of climate change, the 21st-century world is facing the greatest global health threat. Current climate conditions and changes projected by the Intergovernmental Panel on Climate Change predict the impact of rising temperatures on human health. Hence, we developed a user-friendly and accessible tool as Shiny app, exploiting the power of interactive data visualization and real-time analytics to provide policymakers and researchers with a comprehensive platform to monitor heat-attributable mortality during the summer months (https://ficlima.shinyapps.io/mace/). In 2024, we expanded from a national estimate to a provincial one, including data from the current year and the historical series since 2018. Consequently, in the summer of 2024, we estimated higher fractions attributable to extreme heat in the province of Soria with 10%, followed by Lleida with 7% and approximately 6% in Zaragoza, Cuenca, Ciudad Real, and Zamora. Compared to previous years, the summer of 2024 showed a lower extreme heat-related mortality, the third-highest estimated fraction in the past years. The next step will be the implementation of heat-related mortality forecasts based on AEMET's municipal temperature predictions. We hope to contribute to the ongoing discourse on climate change mitigation and public health preparedness, not only within Spain but also as a model for regions facing similar climatic challenges globally.

Comunicaciones de póster

Evaluación de RSPRITE en la Detección de Fraude en Ciencias Sociales”

Autor: Antonio Matas-Terrón [1]

Co-autores: Lourdes Aranda Garrido [1]; Pablo D. Franco-Caballero [1]

[1] *Universidad de Málaga*

La necesidad de publicar, las perspectivas de progreso profesional y la influencia de intereses económicos pueden inducir a los científicos a incurrir en conductas fraudulentas. En este entorno, el ámbito de las ciencias sociales no es inmune al fraude, afectando negativamente la citación de trabajos; si bien, el daño más severo es la pérdida de confianza en la ciencia en sí (Ellenberg, 2000). Para enfrentar este problema, se han sugerido soluciones como la revisión por pares, la replicación y las denuncias, aunque ninguno de estos métodos es totalmente eficaz por sí solo.

La práctica de compartir bases de datos se ha popularizado en el ámbito editorial como un medio para verificar los datos cuando es necesario. A pesar de ello, la manipulación de datos sigue siendo factible. La auditoría y el análisis detallado de los trabajos funcionan como mecanismos adicionales que no previenen la publicación de estudios fraudulentos, sino que intervienen posteriormente, lo que inevitablemente afecta las relaciones entre colegas.
Dentro de este panorama, RSPRITE es una librería de R que puede utilizarse como un recurso proactivo para la detección de fraude.

Actualmente los autores de esta comunicación están desarrollando un estudio piloto que tiene como objetivo analizar la eficacia de SPRITE en la identificación de estudios científicos con datos manipulados en comparación con datos auténticos. Para ello, se está llevando a cabo un diseño experimental controlado con una muestra de datasets: datasets control (datos auténticos); datasets de tratamiento (datos manipulados a partir de datos reales).

En todos los casos se están usando datos a partir de escalas Likert y tipo Likert, junto con datos de preguntas sociodemográficas habituales en estudios sociales de tipo encuesta.

Se espera que los resultados de este estudio ofrezcan argumentos para considerar a RSprite como una herramiento útil o no a la hora de identificar el fraude en investigación en estudios que usan las encuentas dentro de las Ciencias Sociales.

lp.edu: un paquete para introducir las técnicas de optimización lineal en estudios universitarios de grado

Autor: Josep Antoni Martin Fernandez [1]

[1] *Universitat de Girona*

La limitación de todo tipo de recursos hace que hoy en día sea más importante que nunca la optimización de su consumo. La programación lineal ofrece herramientas para la optimización de sistemas organizativos definidos mediante variables deterministas. Este tipo de sistemas organizativos son muy habituales en el campo de las ingenierías y la informática. Las librerías para optimización en R junto con la posibilidad de la propia programación de funciones hacen de RStudio un entorno idóneo para introducir estas técnicas en los cursos avanzados de las ingenierías. Un repaso a los paquetes existentes pone de manifiesto que ninguno de ellos está pensado para introducir a los usuarios en la complejidad de estas técnicas. El paquete lp.edu llena esta laguna.

La estructura de las funciones que conforman el paquete lp.edu tiene en cuenta que los problemas de optimización (PL, PE, PQ, PNL) tienen unos elementos básicos comunes en su modelo: la función objetivo y las restricciones. Estos modelos también comparten unos resultados básicos en su resolución: valor de la función objetivo, valor de las variables de decisión y de las holguras. Estos elementos son los que definen la clase de objeto LP.model que debe definir el usuario como paso inicial en su análisis. Una vez definido el objeto las funciones que conforman el paquete permiten practicar los conceptos más básicos: modificaciones del modelo, representación de la región factible para problemas bidimensionales, obtención de tablas símplex, iteraciones en tablas símplex, resolución del modelo, formulación del problema dual asociado, obtención de los resultados numéricos y gráficos del análisis de sensibilidad, y la resolución de problemas de programación entera.

En este trabajo se presenta de manera esquemática las funcionalidades del paquete lp.edu y su aplicación a los problemas de optimización lineal en el entorno RStudio mediante la confección de informes en R Markdown.

R en la lucha contra los incendios forestales: conociendo la evolución de un incendio

Autor: Marta Rodríguez Barreiro[1]

Co-autores: Manuel Antonio Novo Pérez [1]; María José Ginzo Villamayor [1]

[1] *CITMAga*

En el marco del proyecto CUI de la Agencia Gallega de Innovación (CUI) de la Xunta de Galicia, y en colaboración con la empresa Avincis Aviation Spain SA, se han desarrollado una serie de algoritmos destinados a colaborar en la extinción de los incendios forestales y la gestión posterior del territorio. La mayor parte de los algoritmos han sido implementados utilizando el lenguaje R. En esta presentación se mostrará uno de esos algoritmos desarrollados en R señalando las librerías principales utilizadas y destacando los resultados obtenidos.
El algoritmo seleccionado permite conocer la evolución del perímetro de un incendio forestal en tiempo real a partir de las posiciones de descarga de agua de las aeronaves. Además, a partir de los vientos y el modelo digital del terreno se calculan los vectores de desplazamiento del incendio. Este algoritmo permite conocer la evolución del incendio desde su inicio hasta el instante actual de ejecución, y también permite obtener el avance esperado del incendio en el siguiente período de tiempo. Una vez que se conoce este avance, se detectan las poblaciones y las vías de transporte que están en la dirección de propagación del incendio y se devuelve una lista con todos los elementos que están en riesgo. El usuario puede establecer un vector con todas las distancias a las que quiere obtener los elementos en peligro, y el algoritmo devolverá todos los elementos en riesgo categorizados por las distancias introducidas por el usuario.
Para este algoritmo se conecta con la API de AEMET para la obtención de datos meteorológicos en tiempo real. Para la obtención de las descargas de agua de las aeronaves se conecta con una base de datos. También se conecta con la API del IDEE (Infraestructura de Datos Espaciales de España) para obtener la información de las vías de transporte. Además, se ejecuta desde R un programa externo, denominado WindNinja, que permite obtener los vientos orográficos de la zona del incendio.
El software R da gran versatilidad en el desarrollo del algoritmo ya que permite ejecutar programas externos, como WindNinja. Además, permite conectar con diferentes servicios para la obtención de datos (base de datos, Meteogalicia, AEMET, IDEE⋯). Entre las librerías más utilizadas, se podrían destacar httr para la conexión con servicios web y terra y sf para el tratamiento de los datos espaciales.

Una app Shiny para la visualización de los resultados del Registro OHSCAR

Autor: Patricia Fernández del Valle [1]

[1] *Fundación Pública Andaluza Progreso y Salud*

En 2012, se inició el Registro OHSCAR (Out of Hospital Spanish Cadiac Arrest Registry), mediante un proyecto financiado por el Instituto Carlos III. El Registro OHSCAR es un proyecto de carácter científico, centrado en la investigación sobre resultados en salud en la atención a la Parada Cardiaca Extrahospitalaria (PCREH) en España. Su objetivo es mejorar la supervivencia con buen estado neurológico de estos pacientes.
Se constituye con los datos procedentes de la atención prestada por los diferentes Servicios públicos de Emergencias Extrahospitalarias (SEM) a pacientes que han sufrido una PCREH. Desde su inicio, han participado los SEM que atienden a una población de más de 40 millones de habitantes. OHSCAR sigue las recomendaciones internacionales de estandarización en la recogida y análisis de datos y está integrado en el Registro Europeo de Parada Cardiaca Extrahospitalaria (EuReCa). Cada SEM, y la correspondiente administración sanitaria en la que se integra, es la titular y responsable de la custodia legal de los datos de sus pacientes y de velar por el cumplimiento estricto de lo previsto en la legislación vigente en materia de protección de datos y resto de normativa de aplicación, incluyendo las regulaciones y autorizaciones de los correspondientes comités de ética e investigación. La titularidad de la propiedad intelectual de OHSCAR corresponde a los SEM implicados. En diciembre de 2022, el Ministerio de Sanidad otorgó la declaración de registro de interés sanitario al OHSCAR. Hasta el momento, OHSCAR ha realizado informes en periodos específicos, predefinidos. En estos periodos, el registro elabora un informe global, de ámbito nacional, y uno para cada uno de los SEM participantes. Estos informes son distribuidos a cada una de las administraciones implicadas, y publicados en revistas de impacto.

El objetivo es desarrollar una aplicación web que proporcione al usuario una herramienta fácil, de uso amigable, para visualizar los resultados del Registro OHSCAR.

Shiny es un paquete de R que permite crear aplicaciones web interactivas directamente desde R. La aplicación se instala en un servidor de Shiny, pudiendo ser utilizada por cualquiera usuario con acceso a dicho servidor sin necesidad de tener instalado el programa R.
El diseño de la aplicación emplea la estructura básica de la aplicación "Shiny", que separa la aplicación en menús en la barra lateral y, dentro de cada menú, la separa en pestañas en la sección central de la aplicación, para mostrar los principales resultados obtenidos.
La aplicación proporciona una visualización de la incidencia y de las características principales de la atención prestada por cada SEM ante una PCREH, y los resultados en salud, en términos de supervivencia. Permite explorar subgrupos de variables y subgrupos de pacientes de interés desde el punto de vista clínico, así como la información a nivel nacional y para cada uno de los periodos estudiados. Además, la aplicación incluye los indicadores definidos por la ESCAV.
La aplicación incluye mensajes, observaciones y notas, para facilitar la interpretación de los resultados al usuario menos familiarizado con el proyecto, así como las publicaciones vinculadas al proyecto y un enlace a su página web.

El resultado final es una aplicación que permite al usuario visualizar la información disponible del Resgitro OHSCAR. Las herramientas que el software R pone a nuestra disposición permite la difusión de los proyectos de investigación de cualquier alcance, de una manera ágil y accesible.

phyloshapeR: plotting phylogenies to look like maps

Autor: Ignacio Ramos-Gutiérrez [1]

[1] *Universidad de Sevilla*

The R package phyloshapeR aims to create visual plots of phylogenies to match any shape, including polygons, maps, or any custom shape. It can be already installed from GitHub, and includes an extended tutorial of use (https://iramosgutierrez.github.io/phyloshapeR/).
The workflow of 'phyloshapeR' is to modify a phylogeny' s tree branch lengths in order for they to end matching a shapefile silhouette. To do so, it depends on two extensively used R packages, namely 'ape' (to work with phylogenies) and 'terra' (to calculate distances to a shapefile' s silhouette). Different parameters can be edited within phyloshapeR' s functions, as can be the branch elongation method, the filling depth of internal branches, the number of tips, the coordinates where the root of the phylogeny should be plotted or the plotting options (as colour or width) of the phylogeny.
As a warning to all potential users, the objective of this package is just to create aesthetic phylogenies, never to return data to be used in analyses. However, its outcomes can result visually attractive and may be used for a variety of purposes such as rendering project logos, images for conferences or book covers, combining users' phylogenies and contours for each individual situation.

R en la Evaluación de Estrategias de Gestión de Stocks de Recursos Marinos

Autor: Diana María González Troncoso [1]

Co-autores: Lucía Rueda Ramírez [1]; Margarita María Rincón Hidalgo [1]; María Grazia Pennino [1]; María Soto Ruíz [1]

[1] *Instituto Español de Oceanografía (IEO-CSIC)*

R es la herramienta principal que se emplea actualmente para la evaluación y gestión de stocks de recursos marinos, a través de diferentes paquetes específicos desarrollados en los últimos años y distribuidos de forma abierta y colaborativa en plataformas. Estudios sobre buenas prácticas en evaluación y gestión de stocks pesqueros recomiendan la utilización de estas plataformas y paquetes de R para contribuir a la transparencia, reproducibilidad, homogeneidad, democratización en el uso de aplicaciones y revisión por pares de trabajos sobre evaluación y gestión. Los resultados de estos trabajos tienen implicaciones directas en la sociedad, ya que su objetivo es la explotación sostenible de los recursos pesqueros, los cuales suponen una importante fuente de alimento a nivel mundial. En particular, la Evaluación de Estrategias de Gestión es una disciplina dentro del ámbito de la evaluación y gestión de stocks pesqueros relativamente reciente y en continuo desarrollo y que constituye una de las mejores herramientas hasta el momento para la toma de decisiones de gestión. Esto se debe a que permite anticipar los resultados de las medidas de gestión bajo un entorno de simulación, es decir, permite asesorar a los gestores sobre qué medidas se deben aplicar en el futuro teniendo en cuenta las incertidumbres que afectan al sistema pesquero. Para ello, entre todas las librerías desarrolladas en R, analizamos tres de las más usadas actualmente: FLR_MSE, FLBEIA y OpenMSE. Dado que la aplicación de Evaluación de Estrategias de Gestión requiere de un profundo conocimiento tanto de las pesquerías como del manejo y programación en R, este trabajo pretende servir de utilidad para dotar de elementos de decisión a los posibles usuarios de estas herra mientas evaluando los paquetes según diferentes criterios como son: su facilidad de manejo, facilidad de ejecución de los modelos, dificultad en la incorporación de los datos de entrada, capacidad para implementar nuevas funciones por el usuario, posibilidad de ejecutar evaluaciones para stocks pobres en datos, incorporación de variables ambientales y, finalmente, documentación existente al respecto.

Igualdad de Género en la Transferencia de Conocimiento: Análisis Avanzado con R

Autor: Matilde Pulido Prior [1]

Co-autores: Manuel Fernández Esquinas [2]; María Isabel Sánchez Rodríguez [1]; Olga Salido Cortés [3]

[1] *Universidad de Córdoba*

[2] *IESA CSIC*

[3] *Universidad Complutense de Madrid*

Este estudio examina las desigualdades de género en la transferencia de conocimiento (TC) dentro del ámbito académico, utilizando datos detallados de la convocatoria piloto del sexenio de transferencia (2018-2021) de la ANECA. Empleamos R para desarrollar un enfoque multidisciplinario que combina técnicas de análisis multivariante y Análisis Cualitativo Comparado (QCA) con el objetivo de examinar cómo las disparidades de género afectan la eficacia de la TC. Nuestros hallazgos revelan patrones complejos de participación desigual, ofreciendo perspectivas valiosas para una comprensión más equitativa de la TC.
Nuestro estudio profundiza en los factores que afectan la igualdad de género. Emplear técnicas avanzadas de análisis de datos para identificar las variables más significativas y utiliza el Análisis Cualitativo Comparado (QCA) para descubrir combinaciones de factores que llevan al éxito. Este análisis detallado no solo revela las interacciones complejas entre estas variables, sino que también señala oportunidades claras para intervenir y mejorar la eficacia de la transferencia de conocimiento. La elección de R como herramienta analítica es estratégica, dado que su flexibilidad y capacidad para integrar diversos métodos analíticos facilitan un enfoque comprensivo. Esto facilita abordar la multidimensionalidad de la TC y las cuestiones de género asociadas. La capacidad de R para manejar grandes volúmenes de datos y realizar análisis complejos nos permite no solo obtener una visión detallada de las tendencias observadas, sino también avanzar en nuestra comprensión científica, lo que contribuye al desarrollo de estrategias efectivas para una mayor inclusión en la TC.

Palabras clave: Género y transferencia de conocimiento, Análisis multivariante en R, Desigualdad de género en academia, QCA y eficacia en TC.

La implicación de la mujer en la innovación y la transferencia de conocimiento a través del análisis de la generación de patentes

Autor: María Isabel Sánchez-Rodríguez [1]

Co-autores: Javier Etxabe Oria [2]; Manuel Fernández Esquinas [3]; Matilde Pulido Prior [1]

[1] *Universidad de Córdoba*

[2] *Vicepresidencia Adjunta de Transferencia del CSIC*

[3] *IESA-CSIC*

A través de la transferencia de conocimiento (TC) las universidades y los organismos públicos de investigación trasladan recursos a la empresa, a la administración y a la sociedad civil. A la vez, la TC alimenta la investigación básica de calidad. Por ello, las políticas de innovación tratan de fomentar la TC, aunque se trata de un asunto difícil de conocer y gestionar. Un hecho persistente en la TC es la estratificación entre colectivos de la comunidad académica: entre áreas, perfiles investigadores y tipos de organismos. Además, una de las desigualdades fundamentales se encuentra en el género. Así, este trabajo aborda el grado de implicación de la mujer en la TC a través de la generación de patentes.
La principal fuente de análisis es la base de datos (BD) de patentes del CSIC. El CSIC es la mayor institución española (pública o privada) solicitante de patentes por lo que la BD contiene las solicitudes de prioridad relativas a más de 3.000 patentes y más de 14.000 personas, observadas ininterrumpidamente desde el año 1.994 hasta la actualidad. Estas patentes corresponden a sectores técnicos de

desarrollo y de aplicación muy diversos, incluidos sectores tan emergentes como la nanotecnología. Se dispone de los datos de la patente, de personas y de instituciones participantes, es decir, de numerosas variables de distinto tipo que posibilitan la aplicación de diferentes técnicas de modelización estadística.
En particular, se hará uso del lenguaje R para, en primer lugar, estimar modelos de regresión que permitan determinar diferencias en el porcentaje de participación de la mujer en una patente de acuerdo con aspectos tales como el nivel profesional, el grado de multidisciplinariedad, el área de investigación o el tipo de entidad, entre otros. Posteriormente, se aplicará también análisis cualitativo comparado (QCA) para determinar cómo afectan, no sólo a nivel individual o marginal, sino también combinaciones de las variables independientes en la dependiente, consiguiendo una visión complementaria a la regresión de las relaciones causa-efecto existentes entre las variables.
Como conclusión, este estudio permitirá obtener un mapa representativo de la TC en función de instituciones y especialidades, desarrollando mediciones de los impactos específicos de las condiciones laborales en hombres y mujeres.

Palabras clave: transferencia de conocimiento, patentes, género, R, regresión, análisis cualitativo comparado.

Nuevos retos en los modelos desarrollados en R para la evaluación y gestión de pesquerías bajo incertidumbre

Autor: María Soto Ruiz [1]

[1] *Instituto Español de Oceanografía - CSIC*

La evaluación de pesquerías es el proceso por el cual se obtienen una serie de puntos de referencia a partir de la información de las poblaciones de peces y las flotas que los capturan y que nos informan de si dichas poblaciones se encuentran en buenas condiciones para que su explotación sea sostenible. Es decir, nos permite modelar si estamos extrayendo en las capturas lo que la población regenera cada año. La obtención de estos puntos de referencia y las condiciones pasadas y presentes de explotación relativa a ellos se realiza a través de modelos matemáticos, muchos de los cuales están programados dentro de paquetes de R específicos. La evaluación de pesquerías se puede representar en dos dimensiones que evolucionan en el mismo sentido: Por un lado, según los datos disponibles y, por otro lado, según la complejidad de los modelos matemáticos, que aumenta a medida que se dispone de mayor información. El desarrollo de paquetes de R para la evaluación de pesquerías abarca todos los niveles de complejidad adaptándose a la información disponible. Desde los modelos más sencillos que se emplean en evaluación de stocks pobres en datos a los modelos más complejos integrados que incorporan conjuntamente información sobre la biología, la pesquería, información ambiental al nivel espacial más desagregado posible.
Los principales retos actuales en métodos de evaluación se enfocan hacía: 1) la evaluación de detalles técnicos de modelos concretos de evaluación; 2) la aplicación y distribución de nuevas técnicas que se emplean en evaluación de stocks; 3) la validación de métodos existentes y 4) desarrollo de nuevos métodos. El desarrollo de nuevos modelos requiere la colaboración multidisciplinar que combine el conocimiento matemático y estadístico en algoritmos eficientes de estimación, en programación, en modelado de las hipótesis biológicas y de la dinámica de las poblaciones, en funciones de verosimilitud y en el conocimiento de la gestión de pesquerías. El desarrollo de métodos adecuados en R que sintetice este conocimiento multidisciplinar es clave para avanzar en el complejo proceso de evaluación y gestión de las pesquerías en un entorno de gran incertidumbre. En este trabajo, se realiza una revisión de los principales paquetes de R empleados en pesquerías según los diferentes niveles de información como son los que incluye la librería FLR (FLLife, a4adiags, FLBRP, etc); datalimited, TropFishR, VMStools, SPiCT o JABBA, así como algunos que no son propiamente de pesquerías pero de uso fundamental en evaluación como paquetes estadísticos glmmTMB, lsmeans, mgcv, sdmTMB o espaciales como INLA, VAST o sdmTMB. El objetivo es 1) exponer los distintos niveles de complejidad e incertidumbre existente en la evaluación de pesquerías; 2) revisar las herramientas actuales disponibles en R para la evaluación para en función del grado de información disponible y 3) identificar los retos donde actualmente se está profundizando en el desarrollo de R para la mejora de la evaluación de pesquerías.

Desarrollo de un paquete en R para la implementación de operadores SUOWA y Semi-SUOWA

Autor: Teresa Gonzalez Arteaga [1]

Co-autores: Bonifacio Llamazares Rodríguez [1]; Rocío de Andrés Calle [2]

[1] *Universidad de Valladolid*

[2] *Universidad de Salamanca*

Las medias ponderadas y los operadores OWA (ordered weighted averaging operators) son dos familias de funciones ampliamente utilizadas en la toma de decisiones. A su vez, ambos conceptos son casos especiales de la integral de Choquet, lo que indica su relevancia en el ámbito de la teoría de la decisión.
En la literatura especializada, se han propuesto diferentes procedimientos para generalizar simultáneamente estas dos familias de funciones y crear familias de integrales de Choquet que las incluyan como casos particulares. Entre estas propuestas, destacan los operadores SUOWA y Semi-SUOWA, que mantienen las características distintivas de las medias ponderadas y los operadores OWA, como la capacidad de ponderar valores y la posibilidad de descartar valores extremos.
Aunque hay varios paquetes en R que simplifican el uso de la integral discreta de Choquet, como Kappalab y Rfmtool, estos no abordan de manera específica las necesidades de los operadores SUOWA y Semi-SUOWA. En este trabajo presentamos un nuevo paquete de R denominado WEMOWA, desarrollado con el objetivo de cubrir este vacío y proporcionar herramientas robustas y especializadas para el manejo de estos operadores. Además de incluir funcionalidades avanzadas para SUOWA y Semi-SUOWA, el paquete pone un énfasis particular en las medias ponderadas winsorizadas, ya que estas representan un caso de especial relevancia y aplicabilidad dentro del conjunto de operadores SUOWA. Con WEMOWA buscamos no solo facilitar el análisis y la implementación de estos operadores, sino también promover su uso en aplicaciones prácticas y estudios avanzados.