# XII XORNADA DE USUARIOS DE R EN GALICIA

16 de outubro de 2025

# LIBRO DE RESUMOS

> ORGANIZA

> PATRONICAN

XII XORNADA DE USUARIOS DE R EN GALICIA

# PROGRAMA E RESUMOS

16 de outubro de 2025

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

XII XORNADA DE USUARIOS DE R EN GALICIA

**Editado por: María José Ginzo Villamayor**

**en Santiago de Compostela**

**ISBN: XXX-XX-XX-XXXXX-X**

# Presentación



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

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

O programa contempla vintetrés relatorios ao longo de todo o día. Dos cales oito son convidados e ás outras quince atenderon á chamada de recepción de propostas.

O evento contará con participantes destacados do Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga), da Xunta de Galicia, así como do Instituto Galego de Estatística e das tres universidades galegas ou da Universidad Miguel Hernández de Elche. Tamén participarán académicos de universidades internacionais, como a University of Helsinki, Universidade Federal Fluminense (Brasil) e da Academia da Força Aérea (Brasil), así como das entidades Biostachech ou 3edata.

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

Santiago de Compostela, outubro de 2025

O Comité Organizador

## Comité organizador

María José Ginzo Villamayor
*Universidade de Santiago de Compostela*

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

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

## Comité científico

María José Ginzo Villamayor
*Universidade de Santiago de Compostela*

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

# Información xeral

XII XORNADA DE USUARIOS DE R EN GALICIA

## Data

16 de outubro de 2025

## Lugar de celebración

Salón de Graos. Facultade de Matemáticas (USC)

## Web das xornadas

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

## Certificados

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

# Programa

XII XORNADA DE USUARIOS DE R EN GALICIA

## 16 de outubro de 2025

VERSIÓN PRELIMINAR (suxeita a cambios)

Hora Actividade / Expositor

08:45 - 09:00 Recepción de participantes

09:00 - 09:15 Sesión de apertura: Alberto Cabada Fernández – Decano da Facultade Matemáticas, Daniel Cao Labora – Director do Departamento de Estatística, Mª José Ginzo Villamayor – Comité Científico, Universidade de Santiago de Compostela

09:20 - 09:40 Genómica de la conservación de la especie Daphne gnidium en poblaciones costeras atlánticas de Galicia. Carlos Mirás Viña, Adrián Casanova, Fernando Cabana, Miguel Hermida, Andrés Blanco, Javier Ferreiro, Pablo Ramil, Carmen Bouza, Manuel Vera – Universidade de Santiago de Compostela

09:40 - 10:00 Estimación de filamentos no plano mediante o uso da envoltura r-convexa. Aplicación a datos LiDAR. Héctor González Vázquez, Beatriz Pateiro López, Alberto Rodríguez Casal – Universidade de Santiago de Compostela, CITMAga

10:00 - 10:20 Máis ca un mapa: solución semi-automatizada para xerar cartografía de hábitats. Boris Hinojo Sánchez, Yago Alonso, Federico Cheda, Marco Rubinos – 3edata

10:20 - 10:40 O manexo de imaxes termográficas con R: identificando o perímetro dun incendio forestal. Marta Rodríguez Barreiro, Mª José Ginzo Villamayor – CITMAga

10:40 - 11:00 Análisis funcional y modelado predictivo de energía fotovoltaica con R: un estudio en la UDC. Marina Sánchez Villaverde, Manuel Oviedo de la Fuente, Carlos José Escudero Cascón – Universidade da Coruña

11:00 - 11:20 Web scraping de páxinas dinámicas. Automatizando o buscador con RSelenium. Martín García Cebeiro – Universidade de Santiago de Compostela

**11:20 - 12:00 PAUSA CAFÉ**

12:00 - 12:20 Uso de R en Saúde Pública. Miguel Ángel Rodríguez Muíños, Gael Naveira Barbeito – Consellería de Sanidade

12:20 - 12:40 Da abundancia á evidencia: análise diferencial de datos ómicos en R. Julia García Currás – Biostatech, Universidade da Coruña, CITIC

12:40 - 13:00 HULA R! Potencialidades de R na investigación sanitaria. Adrián Casanova Chiclana, Andrés Blanco, Inés Sambade, Rocío Valín, Pilar Rodríguez Ledo – Unidade de Investigación Clínica de Lugo (HULA, FIDIS)

13:00 - 13:20 Estimación robusta de rexións de referencia condicionais. Aplicación na investigación da diabetes. Sara Rodríguez Pastoriza, Javier Roca Pardiñas, Oscar Lado Baleato – Universidade de Vigo

13:20 - 13:40 Mecanismos post-transcripcionales en cistinosis nefropática: integración transcriptómica y proteómica. Diana Carolina Castro Fernández – Universidade de Santiago de Compostela

13:40 - 14:00 Análise do efecto da consulta electrónica no Servizo de Cardioloxía da Área Sanitaria de Santiago de Compostela e Barbanza. Iria Iglesias Gende, Mercedes Conde Amboage, Francisco Reyes Santías – Universidade de Santiago de Compostela

14:00 - 16:00 PAUSA XANTAR

16:00 - 16:20 Estudio de la comunidad de diatomeas en el río Lor. David Olañeta Suárez, C. Gómez Rodríguez, L. Manel – Universidade de Santiago de Compostela

16:20 - 16:40 Uso de R na Estimación en Áreas Pequenas: Un caso real con datos bidimensionais. María Bugallo Porto, Esteban Cabello, María Dolores Esteban Lefler, Domingo Morales Gonzáez, Agustín Pérez, Sofía Rodríguez Ballesteros – Universidad Miguel Hernández de Elche

16:40 - 17:00 Simulación de escenarios de mercado mediante modelado estocástico en R. Daniel Grande Fernández, Antón Seijo Barro – Universidade de Santiago de Compostela

17:00 - 17:20 O impacto da Obriga de Desembarque da UE nas estratexias de pesca da frota de arrastre costeira española. Santiago Arce Pilo, Adriana Nogueira Gassent, José Antonio Castro Pampillón – Universidade de Vigo

17:20 - 17:40 O uso da cor na difusión de datos estatísticos: unha cuestión nada trivial. Jordán Soubrier Benavides – Instituto Galego de Estatística

**17:40 - 18:00 PAUSA**

18:00 - 18:20 Exemplo de análise de datos na Xunta de Galicia: análise de uso de licencias de software ofimático. Marcos Fernández Arias – Xunta de Galicia

18:20 - 18:40 Aprendendo estatística con LEARNINGSTATS. Elena Fernández Sedes, María Isabel Borrajo, Mercedes Conde Amboage – Universidade de Santiago de Compostela

18:40 - 19:00 R como herramienta para el estudio de patrones biogeográficos. Ramiro Martín Devasa, Sara Martínez Santalla, Carola Gómez Rodríguez, Rosa M. Crujeiras Casais, Andrés Baselga Fraga – University of Helsinki

19:00 - 19:20 Traballar con datos de xeito limpo e eficiente coa libraría tidyverse. Mª José Ginzo Villamayor – Universidade de Santiago de Compostela, CITMAga

19:20 - 19:40 Modelagem para otimizaçao de escalas de fiscais de prova. Luciane Ferreira Alcoforado, João Paulo Martins dos Santos – Academia da Força Aérea (Brasil)

19:40 - 20:00 Capturando Bordas de Grafismos Marajoaras para Construçao de Mandalas Matemáticas. João Paulo Martins dos Santos, Luciane Ferreira Alcoforado – Academia da Força Aérea (Brasil)

Índice

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# O impacto da Obriga de Desembarque da UE nas estratexias de pesca da frota de arrastre costeira española

Santiago Arce Pilo[1], Adriana Nogueira Gassent[2] e José Antonio Castro Pampillón[3]

[1]Universidade de Vigo
[2]Insituto Español de Oceanografía
[3]Insituto Español de Oceanografía

## RESUMO

A Obrigación de Desembarco, introducida pola Unión Europea en 2013 e plenamente vixente desde 2019 para todas as especies suxeitas a regulación, prohibe botar ao mar as capturas non desexadas, esixindo o seu desembarco en porto baixo unha categoría específica que exclúe a súa comercialización para consumo humano. Esta medida pretende incentivar unha pesca máis selectiva e sostible. Porén, a súa implementación suscitou interrogantes sobre a súa eficacia real, os cambios operativos inducidos e os factores que determinan o descarte.

Este traballo analiza o impacto desta normativa nas estratexias de pesca da frota de arrastre costeira española, con especial atención á pescada europea (*Merluccius merluccius*), unha especie de elevada importancia comercial e ecolóxica. A análise baséase en datos do programa de mostraxe científica a bordo, recollidos entre 2009 e 2023 en augas do Cantábrico e Noroeste. Aplicouse un enfoque estatístico en dúas partes, mediante modelos aditivos xeneralizados (GAMs). A primeira parte modela a probabilidade de que se produza un descarte e a segunda, a proporción descartada cando este ocorre. Para identificar posibles cambios derivados da normativa, incluíronse termos diferenciados por período (antes e despois da súa implementación). Todo o proceso de análise e modelización realizouse empregando o software **R**, concretamente a libraría **mgcv** (Wood, 2017).

O estudo analiza de forma explícita os factores que inflúen no descarte, salientando a profundidade, a velocidade do buque, a distancia á costa, a estacionalidade e o recrutamento de xuvenís.

Os resultados mostran que, tras a entrada en vigor da normativa, a proporción de capturas descartadas reduciuse significativamente, mentres que a probabilidade de que ocorra un descarte non variou de forma apreciable. Isto suxire que os pescadores modificaron as súas prácticas para minimizar o volume de captura non desexada, probablemente adoptando estratexias máis selectivas.

Este traballo achega evidencia empírica sobre a resposta adaptativa do sector pesqueiro fronte a un marco regulador complexo, e subliña a utilidade da modelización estatística avanzada para comprender como os factores ecolóxicos, técnicos e espaciais condicionan as decisións de descarte. Os resultados contribúen a mellorar o deseño e a avaliación de políticas pesqueiras máis eficaces e sostibles.

**Palabras e frases chave:** Inclúa aquí as palabras e frases chave. Máximo de seis.

## REFERENCIAS

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

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Uso de R na Estimación en Áreas Pequenas: Un caso real con datos bidimensionais

María Bugallo[1], Esteban Cabello[1], María Dolores Esteban[1], Domingo Morales[1], Agustín Pérez[2] e Sofía Rodríguez-Ballesteros[1]

[1]IUI Centro de Investigación Operativa, Universidade Miguel Hernández de Elche, Alicante
[2]Dpto. de Estudos Económicos e Financeiros, Universidade Miguel Hernández de Elche, Alicante

## RESUMO

A Estimación en Áreas Pequenas combina información auxiliar e modelos estatísticos para obter resultados fiables en dominios con pouca mostra. R ofrece ferramentas específicas para axustar modelos mixtos, predecir indicadores e estimar o seu erro cadrático medio. A aplicación a datos reais céntrase no custo laboral e salarial, por traballador e hora efectiva, en empresas españolas, con dominios definidos por Comunidade Autónoma, división CNAE a dous díxitos e tamaño empresarial. Empregan modelos bivariantes de erro anidado sobre transformacións logarítmicas. A partir deles obtéñense os mellores predictores empíricos, versións conformes e estimacións do erro mediante bootstrap paramétrico. O estudo está financiado polo Instituto Nacional de Estatística no marco das súas iniciativas de apoio á investigación en Estatística Oficial.

**Palabras e frases chave**: estimación en áreas pequenas, estatística oficial, modelos mixtos, datos bidimensionais, custo laboral por traballador, custo salarial por hora efectiva.

## 1. INTRODUCIÓN

A Estimación en Áreas Pequenas é unha rama da estatística que busca obter estimacións fiables en dominios xeográficos ou poboacionais con tamaños mostrais pequenos. En contextos onde as enquisas tradicionais non proporcionan datos abondo para certos subgrupos, estas técnicas permiten "tomar forza prestada" doutros dominios ou variables auxiliares mediante modelos estatísticos, mellorando así a precisión dos resultados. A súa aplicación é especialmente relevante na Estatística Oficial, onde é necesario dispoñer de información para a toma de decisións a nivel desagregado. Para realizar estimacións en áreas pequenas é preciso empregar ferramentas estatísticas avanzadas que permitan axustar modelos, obter predicións e avaliar a incerteza asociada aos resultados.
A linguaxe R constitúe unha plataforma especialmente axeitada para este propósito, xa que ofrece múltiples librarías específicas para a implementación de técnicas de estimación en áreas pequenas. Estas técnicas divídense en dúas grandes categorías: os *modelos a nivel de área*, que relacionan estimacións directas baseadas no deseño mostral con covariables específicas de cada dominio, e os *modelos a nivel de unidade*, que empregan directamente as respostas individuais da enquisa como variables obxectivo. Entre as librarías de R máis destacadas atópanse `sae` (Molina e Marhuenda, 2015), que inclúe modelos clásicos como o de Fay-Herriot (1979) para datos agregados e o de Battese, Harter e Fuller (1988) para datos a nivel individual; `emdi` (Kreutzmann et al., 2019), centrada na estimación de indicadores de pobreza e desigualdade; e `hbsae` (Boonstra, 2022), orientada a modelos bayesianos mediante enfoques xerárquicos. Estas ferramentas converten R nun entorno versátil e potente para o desenvolvemento de aplicacións estatísticas neste ámbito.
Neste traballo preséntase unha metodoloxía para estimar indicadores complexos sobre o custo laboral e salarial a partir da Enquisa Trimestral de Custo Laboral (ETCL), publicada polo Instituto Nacional de Estatística (INE). En concreto, abórdase a estimación do custo laboral por traballador

e do custo salarial por hora efectiva, desagregados por actividade económica (división da Clasificación Nacional de Actividades Económicas (CNAE) a dous díxitos), Comunidade Autónoma e tamaño empresarial (1–49, 50–199, 200 e máis traballadores). Para iso, propóñense modelos bivariantes de erro anidado a nivel que permiten predecir conxuntamente pares de variables (custos, número de traballadores e horas efectivas, entre outros). Esta estratexia evita o uso de cocientes sesgados e garante maior coherencia nas estimacións. Dado que o axuste deste tipo de modelos non está dispoñible nas librarías existentes de `R`, desenvolveuse código propio para implementar a metodoloxía proposta (Esteban et al., 2022a, 2022b). A predición e a avaliación da incerteza baséanse en métodos empíricos e técnicas bootstrap, sen depender da distribución do deseño mostral.

## 2. METODOLOXÍA ESTATÍSTICA

Baixo a teoría predictiva de inferencia en poboacións finitas, os vectores censais que conteñen os valores das variables obxectivo considéranse aleatorios e a súa distribución vén dada por un modelo estatístico. Nesta sección xeralízase o modelo lineal mixto de unidade con intercepto aleatorio, tamén coñecido como modelo de erro anidado (*Nested Error Regression Model*, NER), a un contexto bivariado. Este modelo empregarase como base para a construción de predictores dos indicadores de interese en áreas pequenas e o cálculo de seu erro cadrático medio (*Mean Squared Error*, MSE).

### 2.1 MODELO BIVARIANTE DE ERRO ANIDADO

Sexa $U$ unha poboación de tamaño $N$ dividida en $D$ dominios ou áreas $U_1, \ldots, U_D$ de tamaños $N_1, \ldots, N_D$, respectivamente. Sexa $N = \sum_{d=1}^{D} N_d$ o tamaño global da poboación. Sexa $y_{dj} = (y_{dj1}, y_{dj2})'$ un vector de variables continuas medidas na unidade poboacional $j$ do dominio $d$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. Para $k = 1, 2$, sexa $x_{djk} = (x_{djk1}, \ldots, x_{djkp_k})$ un vector fila que contén $p_k$ variables explicativas e sexa $X_{dj} = \mathrm{diag}(x_{dj1}, x_{dj2})_{2\times p}$ con $p = p_1 + p_2$. Sexa $\beta_k$ un vector columna de tamaño $p_k$ que contén os parámetros de regresión e sexa $\beta = (\beta_1', \beta_2')'_{p\times 1}$. O modelo bivariante de erro aninado (BNER) (Esteban et al., 2022a, 2022b) asume que

$$y_{dj} = X_{dj}\beta + u_d + e_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{1}$$

onde os vectores de efectos aleatorios $\{u_d\}$ e erros aleatorios $\{e_{dj}\}$ son independentes con distribucións $u_d \sim N_2(0, V_{ud})$ e $e_{dj} \sim N_2(0, V_{edj})$, $V_{ud}, V_{edj} \in \mathcal{M}_{2\times 2}$. Na práctica, obsérvase unha mostra $s \subset U$, descomposta en subconxuntos $s_d \subset U_d$, e axústase o modelo (1) a estes datos, utilizando as variables transformadas en escala logarítmica para corrixir asimetrías e discrepancias de escala.

### 2.2 MELLOR PREDICIÓN EMPÍRICA E ESTIMACIÓN DO ERRO

Primeiro proponse predecir parámetros aditivos de dominio que poden expresarse como

$$\delta_d = \frac{1}{N_d}\sum_{j=1}^{N_d} h(y_{dj}), \quad d = 1, \ldots, D, \tag{2}$$

onde $h$ é unha función medible coñecida $\mathbb{R}^2 \to \mathbb{R}$. En concreto, as medias poboacionais das transformacións logarítmicas inversas calcúlanse considerando $h(y_{dj}) = z_{dj1} = \exp(y_{dj1}) - c_1$ e $h(y_{dj}) = z_{dj2} = \exp(y_{dj2}) - c_2$, respectivamente, con $c_1, c_2 > 0$. Deste xeito, conséguese que

$$y_{dj1} = \log(z_{dj1} + c_1) \quad \text{e} \quad y_{dj2} = \log(z_{dj2} + c_2), \quad c_1, c_2 > 0.$$

Os correspondentes parámetros aditivos de dominio (parámetros $\delta_d$) son

$$\overline{Z}_{d1} = \frac{1}{N_d}\sum_{j=1}^{N_d} z_{dj1}, \quad \text{e} \quad \overline{Z}_{d2} = \frac{1}{N_d}\sum_{j=1}^{N_d} z_{dj2}, \quad d = 1, \ldots, D. \tag{3}$$

O mellor preditor (*Best Predictor*, BP) de $\delta_d$ é $\hat{\delta}_d^{\mathrm{BP}} = E_{y_r}\left[\frac{1}{N_d}\sum_{j=1}^{N_d} h(y_{dj}) \mid y_s\right]$, que depende do vector de parámetros descoñecidos $\psi = (\beta', \theta')'$. Sexa $\hat{\psi} = (\hat{\beta}', \hat{\theta}')'$ un estimador mostral. As covariables coñécense en todas as unidades da poboación. Logo, o BP empíricio (EBP) de $\delta_d$ é

$$\hat{\delta}_d^{\mathrm{EBP}} = \frac{1}{N_d}\left\{\sum_{j\in s_d} h(y_{dj}) + \sum_{j\in r_d} E_{y_r}\left[h(y_{dj}) \mid y_s; \hat{\psi}\right]\right\}, \quad \text{sendo } r_d = U_d \backslash r_d, \quad d = 1, \ldots, D. \tag{4}$$

Na práctica, apróximase mediante Monte Carlo. A continuación, proponse calcular o EBP da razón $R_d = \frac{\overline{Z}_{d1}}{\overline{Z}_{d2}}$, dun xeito análogo, aínda que máis complexo por ser un parámetro non aditivo.
Finalmente, resólvese un problema de predición conforme multidimensional con restricións de positividade, de xeito que o numerador e o denominador da razón sumen os valores prefixados en varios niveis de agregación superiores. Isto garante a coherencia cos resultados publicados polo INE, baseados en metodoloxías máis sinxelas que proporcionan estimacións fiables en áreas agregadas, conferindo así sentido e compatibilidade ás estimacións obtidas a nivel desagregado.
En relación ao cálculo das medidas de erro, as aproximacións analíticas do MSE son difíciles de derivar para a predición de parámetros complexos, como medias poboacionais de transformacións non lineais e cocientes. Por este motivo, proponse empregar o bootstrap paramétrico para estimar o MSE, seguindo a metodoloxía proposta por González-Manteiga et al. (2008).

## 3. APLICACIÓN A DATOS REAIS

A ETCL, elaborada polo INE, publícase con periodicidade trimestral e está deseñada para ofrecer estimacións directas (por exemplo, estimacións de Hájek (1971)) precisas a nivel de Comunidade Autónoma e por grandes sectores da actividade económica. Porén, o seu deseño non permite obter resultados representativos en ámbitos territoriais máis desagregados. Os tamaños mostrais $n_d$ nos cruces considerados oscilan entre 1 e 73, cunha media de 8 e unha mediana de 6. Neste traballo, centramos a atención na modelización conxunta do custo total bruto (CTB) e o número de traballadores (TR), así como do custo salarial (S) e as horas efectivas (HE). A metodoloxía empregada, así como as conclusións xerais sobre a aplicación do modelo BNER a outros cruces, son análogas.
Na Figura 1 preséntase a comparación entre as predicións EBP e as EBP conformes multiplicativas con restricións de positividade do custo laboral por traballador (á esquerda) e do custo salarial por hora efectiva (á dereita), fronte ás estimacións de Hájek (1971), amplamente utilizadas no INE. Obsérvase que as predicións baseadas no modelo BNER suavizan notablemente ás estimacións de Hájek, especialmente nos dominios con tamaños mostrais pequenos. As versións conformes desvíanse lixeiramente ao ter que respectar a suma en niveis agregados superiores.

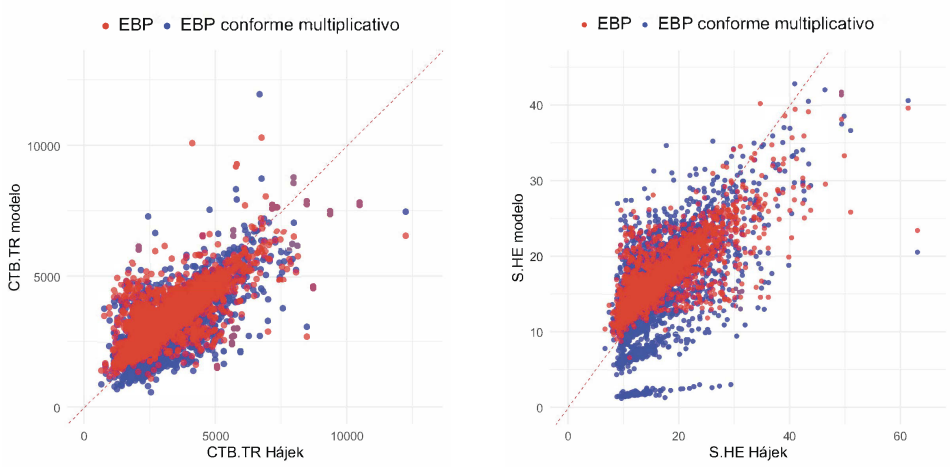



Figura 1: Comparación entre EBP e EBP corrixido conforme multiplicativo do custo laboral por traballador (esquerda) e custo salarial por hora efectiva (dereita), fronte ás estimacións de Hájek.

Na Figura 2 compáranse as estimacións bootstrap da raíz cadrada do MSE (RMSE) do EBP e do EBP conforme multiplicativo con restricións de positividade, segundo os modelos correspondentes, así como a raíz cadrada da varianza directa do estimador de Hájek, para o custo laboral por traballador (á esquerda) e o custo salarial por hora efectiva (á dereita), ordeadas por tamaño mostral. O EBP presenta, en xeral, os menores valores de RMSE, aínda que a súa versión conforme incrementa lixeiramente o erro ao aproximarse máis ao estimador de Hájek. De feito, este último estimador mostra unha variabilidade elevada mesmo en dominios con valores altos de $n_d$ e, nos dominios con $n_d$ máis pequeno, subestima claramente a incerteza, polo que non resulta fiable.

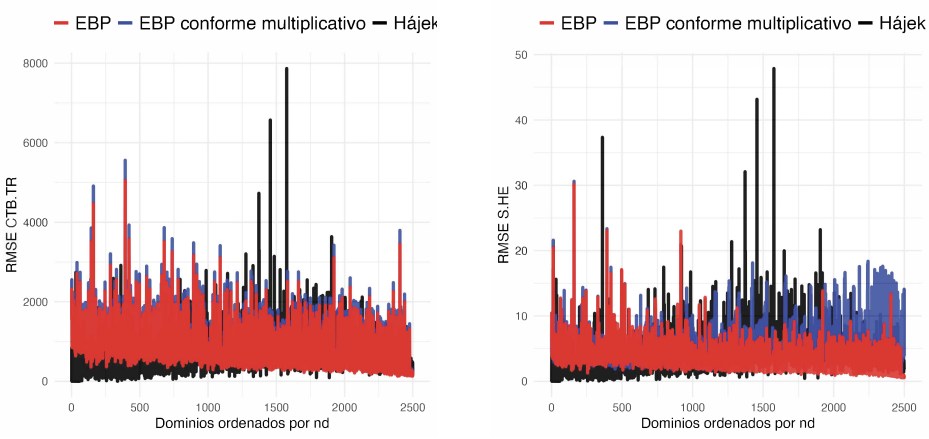



Figura 2: Comparación do RMSE dos preditores segundo o modelo BNER, mediante bootstrap paramétrico, co correspondente á variabilidade das estimacións de Hájek baseada no deseño mostral.

## 4. CONCLUSIÓNS FINAIS

Este traballo aplica o modelo BNER para estimar indicadores complexos da ETCL, como o custo laboral por traballador e o custo salarial por hora efectiva, en áreas pequenas. En concreto, desagregados por Comunidade Autónoma, actividade económica (división CNAE a dous díxitos) e tamaño empresarial. Para garantir coherencia cos resultados agregados publicados polo INE, formulouse un problema de predición conforme multidimensional. A incerteza na predición estimouse mediante bootstrap paramétrico, dada a complexidade de obter expresións analíticas do MSE. Os resultados da aplicación a datos reais melloran a precisión en áreas pequenas do estimador de Hájek, ofrecen máis estabilidade e amosan o potencial desta ferramenta para o seu uso na Estatística Oficial. Toda a metodoloxía foi implementada en `R`, xa que non existen paquetes que permitan axustar este modelo nin aplicar a predición conforme con restricións de positividade como se fixo.

## AGRADECEMENTOS

Esta investigación enmárcase na convocatoria ETD/503/2021 do INE, dentro da liña de investigación 7, centrada na aplicación de técnicas de estimación en áreas pequenas para aproveitar a información auxiliar na desagregación dos resultados das enquisas. Ademais, contou coa colaboración do equipo da Universidade Miguel Hernández de Elche, clave para esta análise.

### Referencias

[1] Battese, G. E., Harter, R. M., Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J Amer Statist Assoc*, **83**(401), 28–36.

[2] Boonstra, H. J. (2022). `hbsae`: Hierarchical bayesian small area estimation. `R` package version 1.2.

[3] Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., Pérez, A. (2022a). Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *J Appl Stat*, **49**(1), 143–168.

[4] Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., Pérez, A. (2022b). Empirical best prediction of small area bivariate parameters. *Scand J Stat*, **49**(4), 1699–1727.

[5] Fay, R. E., Herriot, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J Amer Statist Assoc*, **74**(366a), 269–277.

[6] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., Santamaría, L. (2008). Bootstrap mean squared error of small-area EBLUP. *J Stat Comput Simul*, **78**(4), 443–462.

[7] Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by Basu. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of Statistical Inference*, 236–242.

[8] Kreutzmann, A., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., Tzavidis, N. (2019). `R` package `emdi` for estimating and mapping regionally disaggregated indicators. *J Stat Softwe*, **91**(7), 1–33.

[9] Molina, I., Marhuenda, Y. (2015). `sae`: An `R` package for SAE. *The R Journal*, **7**(1), 81–98.

*XII Xornada de Usuarios de R en Galicia*

*Santiago de Compostela, 16 de outubro do 2025*

# HULA R! Potencialidades de R na investigación sanitaria

Adrián Casanova[1], Andrés Blanco[1], Inés Sambade[1], Rocío Valín[1] e Pilar Rodríguez-Ledo[1]

[1] Unidade de Investigación Clínica de Lugo (UNICL), Hospital Universitario Lucus Augusti (HULA), Fundación Pública Galega Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Lugo, España

## RESUMO

No 2025 arrancou a Unidade de Investigación Clínica de Lugo (UNICL), sita no Hospital Universitario Lucus Augusti (HULA). Este fito está contando coa incorporación de novo material e persoal co obxectivo de catalizar a xeración de coñecemento científico no eido da medicina translacional, buscando a mellora da atención dos pacientes así como da saúde da cidadanía. O traballo actual da Unidade está focalizado no rexistro e análises bioinformáticas de datos clínicos, a divulgación científica e a formación ao persoal asistencial en software libre e gratuíto (e.g., R e JASP). A curto prazo contará con tecnoloxía de secuenciación masiva para a xeración de datos xenómicos (ADN) e transcriptómicos (ARN) que precisarán das mellores ferramentas bioinformáticas. O amplo abano de paquetes de R (> 25.000) o converten nun entorno capaz de traballar desde a estatística descritiva ao aprendizaxe automático asemade dos datos ómicos necesarios para o desempeño desta Unidade. Finalmente, debido a que a meirande parte dos paquetes do entorno R cómpren cos principios FAIR para o software investigador (FAIR4RS; acrónimo do inglés *Findable, Accessible, Interoperable, and Re-usable for Research Software*), o seu emprego resulta coherente cunha aposta firme pola Ciencia Aberta, buscando ser esta Unidade un polo de coñecemento e colaboración.

**Palabras e frases chave:** Saúde, UNICL, secuenciación, principios FAIR4RS , bioinformática, aprendizaxe automática.

## 1. INTRODUCIÓN

A medicina actual enfróntase a un paradoxo: nunca se tiveron tantos datos biolóxicos como na actualidade (*Big Data* biolóxico) e, ao mesmo tempo, nunca foi tan complexo garantir que eses avances cheguen de forma efectiva ao tratamento do paciente. Este espazo é o terreo de xogo da medicina translacional. O seu principal obxectivo é construír sólidas pontes entre a ciencia fundamental e a aplicación clínica, acelerando o ciclo de descubrimento e mellorando de forma tanxible a saúde das persoas. Con este espírito naceu neste 2025 a Unidade de Investigación Clínica de Lugo (UNICL), un proxecto aloxado no Hospital Universitario Lucus Augusti (HULA) coa vocación de ser un polo innovador de coñecemento e colaboracións entre asociacións de pacientes, organismos públicos e privados.

Actualmente, o gran catalizador e o maior desafío para a medicina translacional é a explosión no volume e na complexidade dos datos, tanto estruturados coma non estruturados. Xa non se trata soamente de historias clínicas ou resultados de análises de sangue dun laboratorio. Son datos como as imaxes de resonancias magnéticas en alta resolución ou os procedentes de dispositivos vestibles de monitorización continua (e.g., reloxo intelixente). Ademais, de forma inminente no caso desta Unidade, da xeración a gran escala de datos ómicos. A secuenciación de xenomas (ADN) ou as análises da expresión xénica (ARN) xerarán datos na orde de terabytes. Esta morea de información é unha oportunidade histórica para desenvolver unha medicina máis eficiente, precisa e personalizada, pero só de contar coas ferramentas bioinformáticas máis axeitadas para convertela en coñecemento útil.

Para navegar neste océano de datos non abonda con ter capacidade de computación, precísanse metodoloxías de traballo que garantan que os resultados sexan eficientes, reproducibles e colaborativos. A nosa aposta na UNICL é o software libre e gratuíto e, entre outras opcións, o entorno de programación estatística R. As razóns principais desta elección serían as seguintes: (1) a polivalencia bioinformática grazas ao amplo abano de paquetes aplicables en diferentes campos **{gam}** como os creados, alomenos inicialmente, para campos específicos **{OptimalCutpoints}**; (2) a existencia dunha masa crítica que garante a viabilidade deste proxecto a longo prazo e facilita a existencia de multitude de manuais, foros e titoriais; (3) a dispoñibilidade de R no Centro de Supercomputación de Galicia (CESGA) e finalmente pero non menos importante (4) a súa aliñación cos principios FAIR para o Software de Investigación (FAIR4RS [1]).

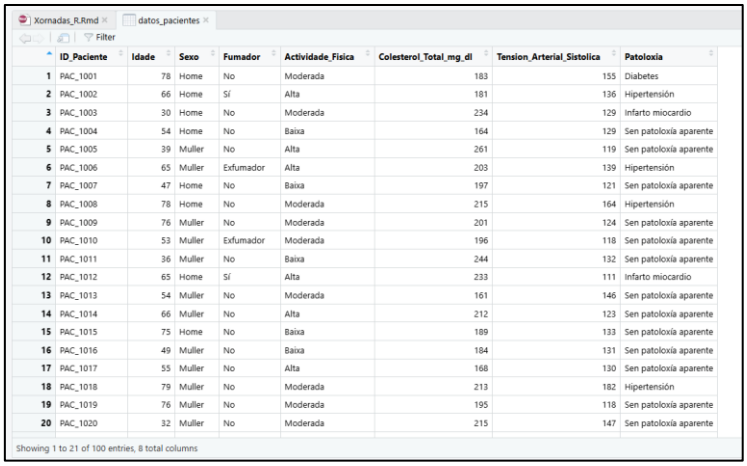
Estes principios recomendan que as ferramentas bioinformáticas sexan: (1) Atopables, grazas a repositorios públicos como **CRAN** (*Comprehensive R Archive Network*); (2) Accesibles, xa que a súa natureza de código aberto e a súa gratuidade elimina barreiras de licenzas privativas ou económicas; (3) Interoperables, pola súa capacidade para comunicarse con outras linguaxes coma Python (e.g., paquete de R **{reticulate}**), manexando ducias de formatos de datos de maneira máis estandarizada; e, fundamentalmente; (4) Reutilizables, porque a filosofía de R promove que unha investigación poida ser o punto de partida para outra.

Ao longo deste documento amosaranse algún usos de diferentes paquetes de R no eido da investigación sanitaria e a súa utilidade para a mellora da atención ao paciente.

## 2. UTILIDADES DE R

A versión de R empregada para a elaboración deste documento foi a 4.5.1 "Great Square Root" traballando nun Entorno de Desenvolvemento Integrado (EDI) proporcionado pola versión de RStudio 2025.09.0+387. Os datos mostrados foron xerados aleatoriamente con Intelixencia artificial (Google Gemini), se ben baseados en situacións realistas.

Un dos primeiros pasos a realizar cos datos, sexan procedentes de análises clínicas, ou o número de lecturas de ADN secuenciadas por mostra (i.e., control de calidade), serían aqueles enmarcados dentro da estatística descritiva mediante a obtención de medidas numéricas como a mediana. Pódese obter esta información a partir dos datos (Figura 1) coas funcionalidades básicas de R **{base}** empregando a función "summary". Con todo, hai paquetes como **{skimr}** e **{gtsummary}** que permiten a obtención de máis información, cun mellor formato (Figura 2) e de forma sinxela: "skim(datos_pacientes)".



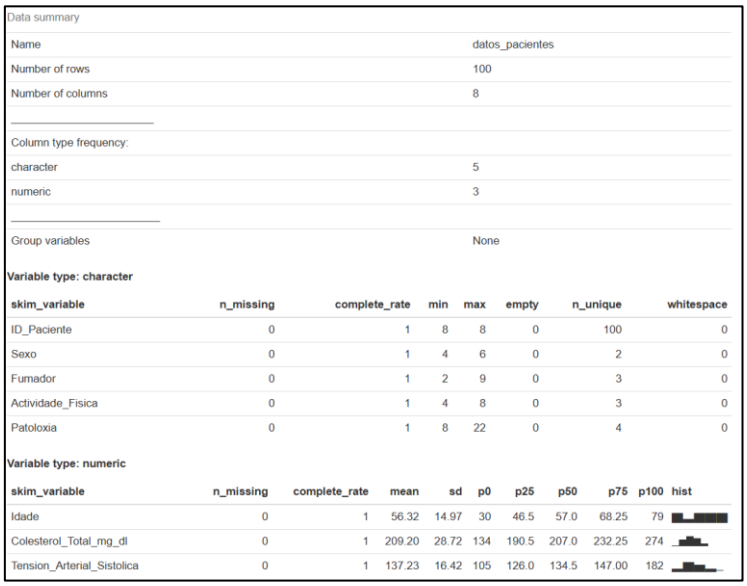
Xornadas_R.Rmd × datos_pacientes ×

Filter

| | ID_Paciente | Idade | Sexo | Fumador | Actividade_Fisica | Colesterol_Total_mg_dl | Tension_Arterial_Sistolica | Patoloxia |
|---|---|---|---|---|---|---|---|---|
| 1 | PAC_1001 | 78 | Home | No | Moderada | 183 | 155 | Diabetes |
| 2 | PAC_1002 | 66 | Home | Sí | Alta | 181 | 136 | Hipertensión |
| 3 | PAC_1003 | 30 | Home | No | Moderada | 234 | 129 | Infarto miocardio |
| 4 | PAC_1004 | 54 | Home | No | Baixa | 164 | 129 | Sen patoloxía aparente |
| 5 | PAC_1005 | 39 | Muller | No | Alta | 261 | 119 | Sen patoloxía aparente |
| 6 | PAC_1006 | 65 | Muller | Exfumador | Alta | 203 | 139 | Hipertensión |
| 7 | PAC_1007 | 47 | Home | No | Baixa | 197 | 121 | Sen patoloxía aparente |
| 8 | PAC_1008 | 78 | Home | No | Moderada | 215 | 164 | Hipertensión |
| 9 | PAC_1009 | 76 | Muller | No | Moderada | 201 | 124 | Sen patoloxía aparente |
| 10 | PAC_1010 | 53 | Muller | Exfumador | Moderada | 196 | 118 | Sen patoloxía aparente |
| 11 | PAC_1011 | 36 | Muller | No | Baixa | 244 | 132 | Sen patoloxía aparente |
| 12 | PAC_1012 | 65 | Home | Sí | Alta | 233 | 111 | Infarto miocardio |
| 13 | PAC_1013 | 54 | Muller | No | Moderada | 161 | 146 | Sen patoloxía aparente |
| 14 | PAC_1014 | 66 | Muller | No | Alta | 212 | 123 | Sen patoloxía aparente |
| 15 | PAC_1015 | 75 | Home | No | Baixa | 189 | 133 | Sen patoloxía aparente |
| 16 | PAC_1016 | 49 | Muller | No | Baixa | 184 | 131 | Sen patoloxía aparente |
| 17 | PAC_1017 | 55 | Muller | No | Alta | 168 | 130 | Sen patoloxía aparente |
| 18 | PAC_1018 | 79 | Muller | No | Moderada | 213 | 182 | Hipertensión |
| 19 | PAC_1019 | 76 | Muller | No | Moderada | 195 | 118 | Sen patoloxía aparente |
| 20 | PAC_1020 | 32 | Muller | No | Moderada | 215 | 147 | Sen patoloxía aparente |

Showing 1 to 21 of 100 entries, 8 total columns

Figura 1: Captura de pantalla do *dataframe* "datos_pacientes" para estatística descritiva.



Data summary

| | |
|---|---|
| Name | datos_pacientes |
| Number of rows | 100 |
| Number of columns | 8 |
| Column type frequency: | |
| character | 5 |
| numeric | 3 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| ID_Paciente | 0 | 1 | 8 | 8 | 0 | 100 | 0 |
| Sexo | 0 | 1 | 4 | 6 | 0 | 2 | 0 |
| Fumador | 0 | 1 | 2 | 9 | 0 | 3 | 0 |
| Actividade_Fisica | 0 | 1 | 4 | 8 | 0 | 3 | 0 |
| Patoloxia | 0 | 1 | 8 | 22 | 0 | 4 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| Idade | 0 | 1 | 56.32 | 14.97 | 30 | 46.5 | 57.0 | 68.25 | 79 | |
| Colesterol_Total_mg_dl | 0 | 1 | 209.20 | 28.72 | 134 | 190.5 | 207.0 | 232.25 | 274 | |
| Tension_Arterial_Sistolica | 0 | 1 | 137.23 | 16.42 | 105 | 126.0 | 134.5 | 147.00 | 182 | |

Figura 2: Captura de pantalla do resultado obtido co paquete **{skimr}**.

Exemplo de paquetes específicos do eido da saúde serían **{OptimalCutpoints}** e **{psych}**. **{OptimalCutpoints}**, serve, entre outras funcionalidades, para determinar o punto de corte óptimo en probas diagnósticas, clasificando os pacientes en sans e enfermos. Se realizamos unha proba piloto cun cuestionario de saúde, pode ser de interese calcular o coeficiente alfa de Cronbach, que mide cos resultados obtidos a consistencia interna das preguntas entre si. O seu rango de valores vai de 0 a 1, sendo un valor más elevado indicativo de que as preguntas están medindo a mesma cualidade de maneira consistente (e.g., ansiedade; Táboa 1). Coa función "alpha" do paquete {**psych}** obtense este índice calculado para o cuestionario global e para cada pregunta, a fiabilidade do test descartando cada unha das preguntas, etcétera.

| **Código Pregunta** | **Durante as dúas últimas semanas, con que frecuencia sentiu molestias por...** |
|---|---|
| PA_1 | Sentirse nervioso/a ou intranquilo/a? |
| PA_2 | Non poder controlar a súa preocupación? |
| PA_3 | Preocuparse demasiado por diferentes cousas? |
| PA_4 | Ter dificultades para relaxarse? |
| PA_5 | Sentir medo coma se algo terrible puidera pasar |

Táboa 1: Exemplo dalgunhas preguntas para testar a ansiedade. Baseado no test de *Mental Health America* (https://screening.mhanational.org/screening-tools/test-de-ansiedad/).

Dentro de R hai multitude de paquetes para aplicar a Aprendizaxe Automática (do inglés: *machine learning*), conseguindo que os ordenadores poidan "aprender" dos datos sen seren programados explicitamente. De forma xeral, poderíanse clasificar as aprendizaxes automáticas en tres tipos: (1) Supervisada, con datos de adestramento (~80%) e validación (~20%) **{caret, randomForest, tidymodels}**; (2) non supervisada, clusterizando as entradas dos datos en base a patróns compartidos sen hipóteses previas **{dbscan, fda.usc}** e (3) por reforzo, recompensando as respostas correctas **{ReinforcementLearning}**. Estas análises e ferramentas teñen utilidades claras no eido sanitario. A modo de exemplo, aproveitando o elevado volume de rexistros sanitarios poderíase, mediante aprendizaxe automática supervisada, estimar a probabilidade de reingreso hospitalario dun paciente nun determinado rango temporal ou a duración destes ingresos. Isto permitiría a previsión e a optimización dos recursos sanitarios.

É habitual que nos diferentes campos de investigación, haxa revisións científicas sobre os paquetes de R dispoñibles e as súas metodoloxías. No campo das ómicas existen varios exemplos coma se ve esquematicamente na Figura 1 do traballo de Chua et al. (2024) [2]. Existen paquetes como **{TCGAbiolinks}** e **{GEOquery}**, sitos no repositorio **Bioconductor**, para a descarga de datos de diferentes bases de datos públicas como *Genomic Data Commons* (GDC; https://portal.gdc.cancer.gov/) e *Gene Expression Omnibus* (GEO; http://www.ncbi.nlm.nih.gov/geo/), respectivamente. Existen paquetes paraugas como **{BGData}** capaces de realizar unha batería de análises xenómicos (e.g., Estudos de Asociación do Xenoma Completo; GWAS) coa mínima memoria RAM, para a anotación funcional das variantes atopadas no ADN **{VariantAnnotation}**, visualización do xenoma **{genoPlotR**, **RIdeogram}**, análises de expresión diferencial de miles de xenes con **{DESeq2}**, análise de enriquecemento funcional con **{gprofiler2}** e un longo etcétera.

## 3. CONCLUSIÓNS

Co entorno de R (case) todas as análises bioinformáticas son posibles, sen custe económico e sen as barreiras das licenzas privativas, cumprindo polo xeral cos principios FAIR4RS. Debido a isto, recomendamos o seu uso en calquera campo de investigación. Consideramos recomendable para aquelas persoas que cambien de campo de traballo a exploración previa das posibilidades de R para o deseño das súas novas rutas bioinformáticas.


## AGRADECEMENTOS

O financiamento para o desenvolvemento da UNICL provén da convocatoria do Instituto de Salud Carlos III baixo o PERTE para la Salud de Vanguardia do Ministerio de Ciencia e Innovación no marco do Plan de Recuperación, Transformación y Resiliencia, regulada mediante la Orden CNU/709/2024, de 1 de julio (BOE nº 164, de 08/07/2024).


## Referencias


[1] Barker M., Chue Hong N.P., Katz D.S., Lamprecht A-L., Martínez-Ortiz C., Psomopoulos F., Harrow J., Castro L.J., Gruenpeter M., Martínez P.A., Honeyman T. (2022). Introducing the FAIR Principles for research software. *Scientific Data* 9, 622. https://doi.org/10.1038/s41597-022-01710-x

[2] Chua E.W., Ooi D.J., Muhammad N.A.N. (2024). A concise guide to essential R packages for analyses of DNA, RNA, and proteins. *Molecules and Cells* 47 (11), 100120. https://doi.org/10.1016/j.mocell.2024.100120


As referencias dos paquetes de R pódense consultar usando este QR

*XII Xornada de Usuarios de R en Galicia*

*Santiago de Compostela, 16 de outubro do 2025*

# Mecanismos post-transcripcionales en cistinosis nefropática: integración transcriptómica y proteómica.

Diana Carolina Castro Fernández[1]

[1] Fundación Instituto de investigación Sanitaria de Santiago

**RESUMO**

La cistinosis, debido a su baja prevalencia, es una enfermedad poco conocida y subdiagnosticada, lo que subraya la importancia de incrementar la concienciación médica y de desarrollar herramientas basadas en biomarcadores para facilitar un diagnóstico temprano y preciso. En este contexto, se propone realizar un análisis integrativo en R de datos de transcriptómica y proteómica, que se encuentran en repositorios publicos, de fibroblastos de pacientes con cistinosis. Este enfoque permitirá identificar las coincidencias y discrepancias entre los niveles de expresión génica (medidos mediante RNA-seq) y la abundancia proteica (analizada mediante LC-MS/MS). Se busca detectar genes y proteínas con correlación directa (alta transcripción/alta expresión) o divergencias significativas (alta transcripción pero baja expresión proteica, y viceversa), lo que podría indicar la existencia de mecanismos de regulación post-transcripcional. Estos resultados contribuirán al entendimiento de los mecanismos moleculares subyacentes a la cistinosis, ofreciendo potenciales biomarcadores y posibles dianas terapéuticas.

**Palabras e frases chave:** Cistinosis, Transcriptomica, Proteomica, Biomarcadores

## 2. MATERIAL Y METODOS

2.1 Obtención de datos:

Proteómica: Se recopilaron un conjunto de datos proteómicos basados en espectrometría de masas provenientes del repositorio publico MassIVE identificados con el código MSV000095020, el conjunto de datos pertenece a fibroblastos primarios de piel no transformados de pacientes con cistinosis (n=12) y controles no afectados (n=7) provenientes del Instituto Coriell para la Investigación Médica (Camden, NJ, EE. UU.).

Transcriptómica: El conjunto de datos de transcriptómica proviene de un perfil de expresión por micromatriz que se descargó del repositorio GEO, identificado con el código GSE190500, el conjunto de datos pertenece a fibroblastos de pacientes con cistinosis (n=8) y controles no afectados (n=6) provenientes del Instituto Coriell para la Investigación Médica (Camden, NJ, EE. UU.).

2.2 Analisis de expresión diferencial:

El análisis se realizó en el entorno RStudio (versión 31.7.7) con R (versión 4.3.3). Para identificar genes y proteínas diferencialmente expresados, se realiza un análisis univariado

aplicando un modelo lineal con estimación moderada de la varianza, mediante el paquete *limma* de Bioconductor.

Se construyó un volcano plot para visualizar de forma conjunta la magnitud del cambio de expresión (*log2FoldChange*) y la significancia estadística (*p*-valor ajustado) de cada gen o proteína. La visualización se realizó en el entorno RStudio (versión 31.7.7) con R (versión 4.3.3), empleando las librerías dplyr, ggplot2, ggrepel, cowplot y readxl.

2.3 Análisis de Redes de Coexpresión Génica Ponderada (WGCNA) multi-bloque:
Para complementar la integración de datos, se aplicó el análisis WGCNA en formato multi-bloque, con el objetivo de identificar módulos de genes y proteínas que compartieran patrones de coexpresión consistentes entre los conjuntos de datos. Se comenzó calculando la matriz de correlación para cada bloque y se determinó el parámetro de umbral de suavizado (soft-thresholding) para construir la red de adyacencia ponderada. Los bloques se integraron mediante la función blockwiseModules del paquete WGCNA, ajustando los parámetros de agrupamiento para cada bloque según su tamaño y estructura. Los módulos identificados se visualizaron mediante mapa de calor (heatmap) para facilitar la identificación de grupos de genes o proteínas con perfiles similares. Los mapas de calor se generaron mediante las librerías pheatmap, utilizando RStudio (versión 31.7.7) con R (versión 4.3.3).

3. RESULTADOS

Análisis de expresión diferencial.

Para ambas representaciones graficas tenemos en el eje de las ordenas las diferencias de expresión de los genes mediante la estimación reducida de logaritmo de Fold Change (LFC), mientras que en el eje de las abscisas se encuentra el -log10 del p-valor. En la proteómica, obtenemos 92 PDE, de las cuales 39 proteínas se encuentran sobreexpresadas y 53 infraexpresadas (Figura 1).

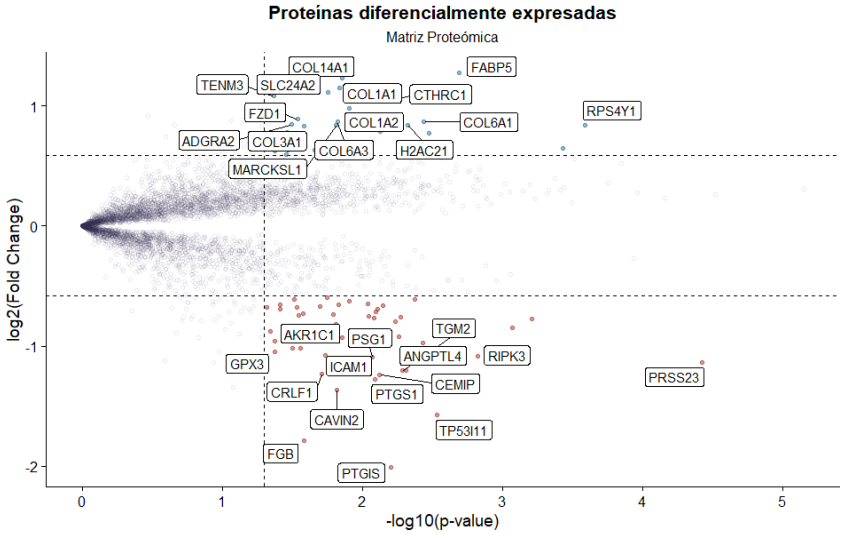



*Figura 1. Volcano plot de genes diferencialmente expresados en la matriz proteómica. El gráfico muestra la distribución de las proteínas en función del cambio de su abundancia (eje Y: $log_2$ Fold Change) y el nivel de significancia estadística (eje X: $-log_{10}$ p-valor). Los puntos en azul representan proteínas significativamente sobreexpresadas, los puntos en rojo corresponden a proteinas infraexpresados en muestras de cistinosis respecto a controles. Las proteínas destacadas con etiquetas corresponden a las que tienen las variaciones más relevantes tanto en magnitud como en significancia*

En la transcriptómica, donde había mayor número de observaciones totales, tenemos 1460 PDE, de las cuales 844 se encuentran sobreexpresadas y 616 infraexpresadas (Figura 2).

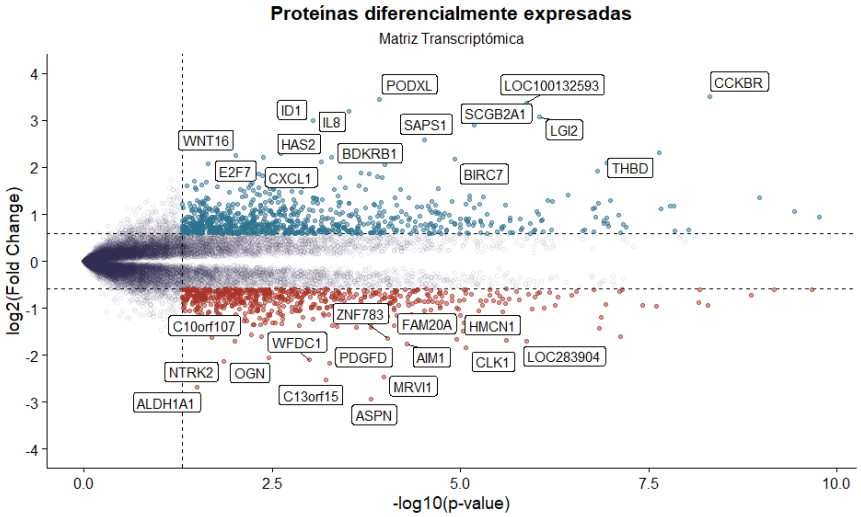



*Figura 2. Volcano plot de genes diferencialmente expresados en la matriz transcriptómica. El gráfico muestra la distribución de los genes en función del cambio de expresión (eje Y: $log_2$ Fold Change) y el nivel de significancia estadística (eje X: $-log_{10}$ p-valor). Los puntos en azul representan genes significativamente sobreexpresados, los puntos en rojo corresponden a genes infraexpresados en muestras de cistinosis respecto a controles. Los genes destacados con etiquetas corresponden a los que presentaron las variaciones más relevantes tanto en magnitud como en significancia*

Análisis de Redes de coexpresión Génica Ponderada (WGCNA) multi-bloque

Este análisis permitió identificar módulos de genes (T_Mn) y de proteínas (P_Mn) con patrones de coexpresión, mediante un agrupamiento jerárquico aplicado a la matriz de similitud topológica. Posteriormente, al evaluar la correlación entre los eigengenes de los módulos transcriptómicos y proteómicos, se observó en el gráfico (Figura 3) la presencia de módulos con correlaciones altamente positivas (valores cercanos a 1, representados en rojo intenso), lo que sugiere una relación directa entre la expresión génica y la abundancia proteica. De manera complementaria, también se detectaron módulos con correlaciones negativas (valores cercanos a –1, en azul intenso), lo que apunta a posibles procesos de regulación postranscripcional o a variaciones en la estabilidad de las proteínas.

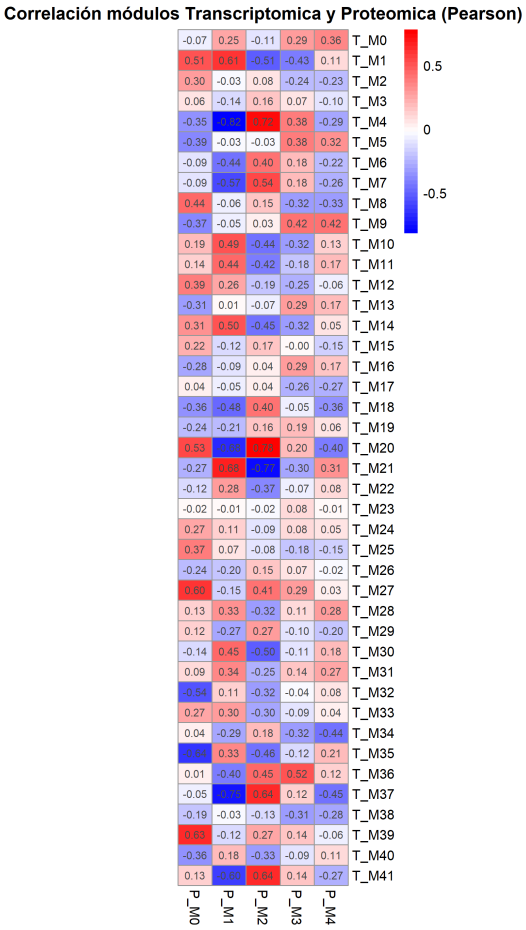



*Figura 3. Correlación entre módulos transcriptómicos y proteómicos identificados mediante WGCNA multi-bloque. Heatmap con las correlaciones de Pearson entre los eigengenes de los módulos transcriptómicos (T_Mn) y proteómicos (P_Mn). Los valores positivos (en rojo) representan asociaciones directas entre expresión génica y abundancia proteica, mientras que los valores negativos (en azul) reflejan posibles procesos de regulación postranscripcional o diferencias en la estabilidad de las proteínas. La intensidad del color indica la magnitud de la correlación, con valores cercanos a ±1 considerados como correlaciones fuertes.*

## Referencias


Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47. https://doi.org/10.1093/nar/gkv007

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (R package version 1.1.4) [Computer software]. https://CRAN.R-project.org/package=dplyr

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Slowikowski, K., Schep, A., Hughes, S., & Lukauskas, S. (2024). *ggrepel: Automatically position non-overlapping text labels with 'ggplot2'* (R package version 0.9.5) [Computer software]. https://CRAN.R-project.org/package=ggrepel

Wilke, C. O. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'* (R package version 1.1.1) [Computer software]. https://CRAN.R-project.org/package=cowplot

Wickham, H., & Bryan, J. (2023). *readxl: Read excel files* (R package version 1.4.3) [Computer software]. https://CRAN.R-project.org/package=readxl

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics, 9*, 559. https://doi.org/10.1186/1471-2105-9-559

Kolde, R. (2019). *pheatmap: Pretty heatmaps* (R package version 1.0.12) [Computer software]. https://CRAN.R-project.org/package=pheatmap

*XII Xornada de Usuarios de R en Galicia*

*Santiago de Compostela, 16 de outubro do 2025*

# Máis ca un mapa: unha solución semi-automatizada para xerar cartografía de hábitats en espazos naturais protexidos

Federico Cheda[1], Yago Alonso[1], Boris Hinojo[1] e Marco Rubinos[1]

[1] 3edata Ingeniería Ambiental S.L.

## RESUMO

A obtención de cartografía para a monitorización e xestión de hábitats en zonas de alto valor ecolóxico, como as integradas na Rede Natura 2000, esixe a manipulación e análise de conxuntos de datos xeoespaciais masivos cunha temporalidade recorrente. Os métodos cartográficos convencionais demostran ser ineficientes fronte a este volume de datos e a necesidade de respostas áxiles. Este traballo introduce unha aproximación semi-automatizada fundamentada no que denominamos 'Modelo de Cartografía de Hábitat', unha implementación integramente desenvolvida en linguaxe R. Este modelo baséase en técnicas de aprendizaxe automático para a segmentación e clasificación de hábitats en tipos do sistema EUNIS e de Interese Comunitario (Directiva de Hábitats) a partir de imaxes de satélite.

A elección da linguaxe R maniféstase na súa capacidade de integración con librarías de intelixencia artificial, xunto coa súa optimizada eficiencia no procesamento distribuído de datos xeoespaciais (*raster* e vectoriais). Esta simbiose permite a execución de fluxos de traballo complexos de forma robusta e a gran escala, superando as limitacións computacionais habituais.

En termos operativos, o modelo permite a identificación e o seguimento de hábitats de interese comunitario de forma coherente e optimizada, abaratando o proceso de xeración de cartografía, e principalmente da súa actualización. O integración do coñecemento local e capas temática previas é crítica para a usabilidade da cartografía por parte dos xestores de espazos naturais e na súa toma de decisións. A solución proposta representa un marco analítico, escalable e replicable para a monitorización e conservación de espazos naturais.

**Palabras e frases chave:** Cartografía de Hábitats, EUNIS, Hábitats de interese comunitario, rede Natura 2000, aprendizaxe automático, R.

## 1. INTRODUCIÓN

A cartografía e monitorización de hábitats en espazos de alto valor ecolóxico, como os integrados na Rede Natura 2000, son fundamentais para cumprir os obxectivos de conservación establecidos pola Directiva de Hábitats da Unión Europea (92/43/CEE) [1]. Non obstante, os métodos cartográficos tradicionais resultan insuficientes para xestionar os grandes volumes de datos necesarios para un seguimento continuo e a grande escala, debido aos seus elevados custos e tempos de produción; ou ben non integran información previa e coñecemento local relativo aos hábitats para xerar unha cartografía útil en termos de xestión dun espazo natural.

Para dar resposta a este desafío, no marco do proxecto de investigación e innovación *Nature FIRST* (Horizonte Europa, acordo 101060954), desenvolveuse o *'Modelo de Cartografía de Hábitats'*. O obxectivo xeral de *Nature FIRST* é reverter a perda de biodiversidade mediante o desenvolvemento de ferramentas que permitan pasar de accións reactivas a estratexias proactivas e preventivas, baseadas en datos. Neste contexto, o modelo preséntase como unha solución semi-automatizada, desenvolvida integramente na linguaxe de programación R, que emprega técnicas de aprendizaxe automático, datos de teledetección e coñecemento local para a clasificación de hábitats.

A ferramenta está deseñada para xerar cartografía de hábitats segundo as clasificacións EUNIS [2], e a súa correspondencia cos Hábitats de Interese Comunitario (HIC), utilizando imaxes de satélite de alta resolución como as do programa Copernicus Sentinel-2 (clasificación baseada en píxeles), así como coas de moi alta resolución (clasificación baseada en obxectos). Deste xeito, búscase reducir custos, mellorar a identificación dos hábitats, acelerar os procesos de actualización e facilitar unha xestión máis eficaz dos espazos naturais protexidos.

## 2. METODOLOXÍA

O *Modelo de Cartografía de Hábitats* estrutúrase como unha arquitectura modular e escalable programada en R, aproveitando a súa capacidade para o procesamento de datos xeoespaciais e a súa integración con librarías de aprendizaxe automática como randomForest. O fluxo de traballo organízase en base a unha cuadrícula de referencia europea (ETRS89-LAEA) de 10x10 km, o que permite adaptar a análise ás particularidades bioxeográficas de cada zona e optimizar o rendemento computacional mediante o procesamento en paralelo.

A metodoloxía divídese en cinco bloques principais de procesos:

**Bloque 1**: Definición do proxecto. Nesta fase inicial, o usuario define a Área de Interese (AoI), selecciona as imaxes multiespectrais que se van empregfar e achega os datos de entrada (p.ex. imaxes Sentinel-2). Estes inclúen capas obrigatorias como a delimitación da área, a grella de referencia e, fundamentalmente, unha rede de áreas de adestramento. Tamén se integran capas temáticas opcionais pero recomendadas (mapas de ríos, estradas, litoloxía, modelos de elevacións, etc.) que enriquecen o modelo e coas que se garda a coherencia espacial e temática.

**Bloque 2**: A xestión de áreas de adestramento, que inclúe a súa creación, actualización e mantemento, que serven para o adestramento do algoritmo *Random Forest* [3]. Para cada cela de 10x10 km, o modelo extrae as estatísticas de reflectancia das imaxes e os índices de vexetación (como o NDVI) correspondentes ás áreas de adestramento. Así mesmo, posibilítase a creación de múltiples clasificadores por cela se se usan máscaras temáticas, como unha que separe a superficie forestal da non forestal.

**Bloque 3**: Clasificación inicial. Unha vez adestrado, o modelo preditor aplícase a todos os píxeles da área de interese, xerando un mapa de clasificación inicial coas clases de hábitats EUNIS adestradas. O sistema conserva o máximo nivel de concreción do sistema xerárquico EUNIS.

**Bloque 4**: Refinamento e Xeneralización. Este é un paso clave para mellorar a calidade da clasificación inicial. Aplícanse unha serie de regras cartográficas, bioxeográficas e semánticas para corrixir erros, engadir clases non identificables espectralmente e asegurar a coherencia con outras fontes de datos oficiais. Neste proceso intégrase coñecemento local relativo aos hábitats a clasificar, que doutra

maneira non se poderían integrar no resultado final. Este proceso organízase en tres estratexias:

◦ Asignación directa de clases a partires de cartografía oficial preexistente (ex. estradas, edificios, ríos).
◦ Regras baseadas na análise contextual de capas temáticas (ex. identificar un bosque de ribeira analizando a vexetación próxima a un curso fluvial).
◦ Regras de decisión baseadas en atributos dimensionais ou de veciñanza sobre as clases xa clasificadas.

Posteriormente, un proceso de xeneralización agrupa polígonos pequenos para simplificar visualmente o mapa, mantendo a trazabilidade da información orixinal.

**Bloque 5**: Estandarización e exportación. O mapa de clases EUNIS final convértese á clasificación de Hábitats de Interese Comunitario (HIC) mediante táboas de correspondencia definidas previamente. O resultado expórtase con metadatos conformes á Directiva INSPIRE, garantindo a súa interoperabilidade.

## 3. RESULTADOS

O *Modelo de Cartografía de Hábitats* foi desenvolvido e probado en diversos espazos naturais protexidos europeos no marco do proxecto *Nature FIRST*. Os sitios de proba inclúen ecosistemas variados que cobren seis rexións bioxeográficas distintas: área de Maramures (Romanía e Ucraína), a montaña de Stara Planina (Bulgaria), o Delta do Danubio (Romanía) e a serra de Ancares-Courel en Galicia (España).

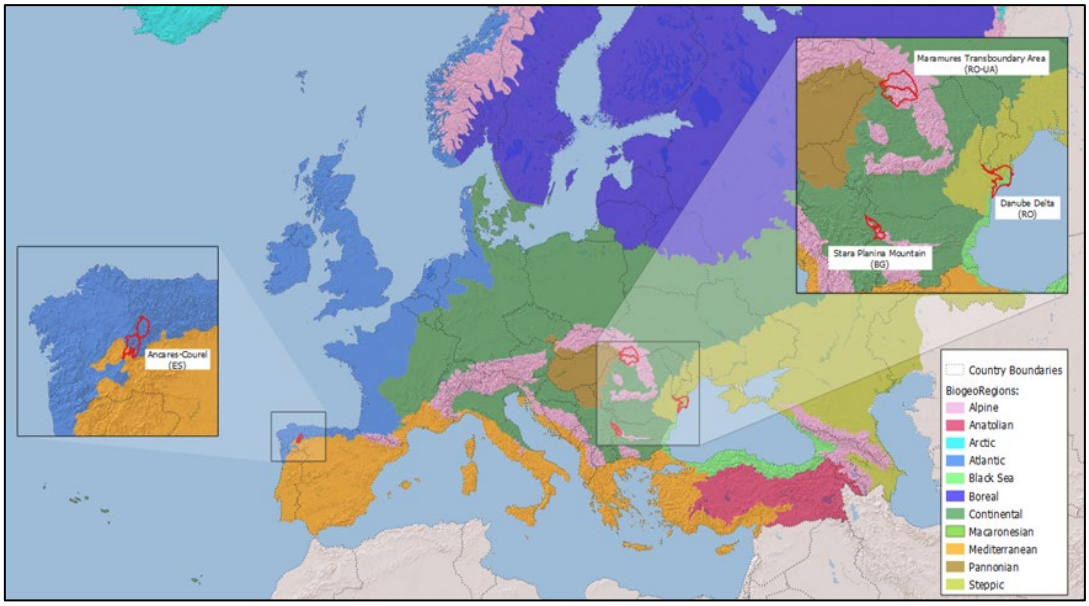


Figura 1: Localización áreas de aplicación de cartografía de hábitats Nature FIRST

O primeiro lugar onde se aplicou e validou o modelo de forma exhaustiva foi a de Ancares-Courel, que abrangue unha superficie de 2.300 km$^2$. Para este espazo, empregouse unha serie temporal de seis imaxes Sentinel-2 e unha rede de áreas de adestramento definida mediante fotointerpretación e traballo de campo.

A precisión do modelo foi avaliada mediante dous métodos independentes: unha validación interna automatizada e unha validación externa manual sobre polígonos aleatorios. Os resultados foron altamente satisfactorios, obtendo un índice Kappa superior a 0,90 na validación interna e superior a 0,80 na validación externa.

Como produto final para Ancares-Courel, o modelo logrou identificar e delimitar cartograficamente 19 hábitats de interese comunitario, dos cales 3 son prioritarios, sobre un total de 37 HIC teoricamente presentes na zona segundo a documentación oficial. Esta diferenza explícase polas limitacións da resolución espacial das imaxes Sentinel-2 para detectar hábitats de superficie moi reducida ou que non son directamente cartografables. O resultado final é un conxunto de produtos cartográficos dixitais que detallan a distribución dos hábitats a diferentes niveis de clasificación (EUNIS e HIC).

## 4. CONCLUSIÓNS

O **Modelo de Cartografía de Hábitats** demostra ser unha alternativa viable, precisa e eficiente para a xeración de cartografía de hábitats en espazos protexidos, superando moitas das limitacións dos métodos convencionais, e posibilitando a compatibilidade con modelos de datos estándar como o do Ministerio de Transición Ecolóxica.

Entre as súas principais vantaxes destacan:

- **Precisión e flexibilidade**: O modelo alcanza niveis de precisión elevados e permite a integración flexible de múltiples fontes de datos xeoespaciais, **harmonizando** os resultados coa cartografía oficial existente.
- **Eficiencia e baixo custo**: A automatización do proceso reduce drasticamente os custos de produción e actualización. A súa replicabilidade facilita a xeración de series temporais para o seguimento da dinámica dos hábitats.
- **Accesibilidade e software libre**: A solución foi desenvolvida sobre tecnoloxías de código aberto (R, GDAL, GRASS GIS) e utiliza fontes de datos de acceso libre (Copernicus), o que garante a súa accesibilidade e reduce os custos de implantación. Ademais, o seu funcionamento non require coñecementos avanzados por parte do usuario final.
- **Adaptabilidade e mellora continua**: A arquitectura modular por celas adáptase ás particularidades de cada territorio, e o sistema de regras permite unha mellora continua do modelo a medida que se incorpora novo coñecemento ou cartografía de referencia.

A principal limitación identificada reside na resolución espacial das imaxes Sentinel-2 (10-60 metros), que dificulta a cartografía de hábitats moi pequenos, fragmentados ou que se atopan en mosaicos complexos con outras coberturas.

En conclusión, o modelo proposto constitúe unha ferramenta potente, escalable e de baixo custo que proporciona aos xestores e investigadores información actualizada e fiable para a monitorización e conservación da biodiversidade, apoiando a toma de decisións informadas e o cumprimento das directivas ambientais europeas.

## AGRADECEMENTOS

Este traballo forma parte do proxecto Horizonte Europa Nature FIRST (acordo 101060954) financiado pola Comisión Europea.

**Referencias**

[1] EEA. (2014). *Terrestrial habitat mapping in Europe: an overview*. EEA Technical report n.1/2014.

[2]. EEA (ETC/BD) (2013). EUNIS Habitat Classification 2012 - a revision of the habitat classification descriptions (https://www.eea.europa.eu/en/datahub).

[3] Belgiu M., Drăguț L. (2016). Random Forest in Remote Sensing: A Review of Applications and Future Directions. Isprs Journal of Photogrammetry and Remote Sensing 114, 24–31.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Exemplo de análise de datos na Xunta de Galicia: análise de uso de licencias de software ofimático

Marcos Fernández Arias[1]

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

**RESUMO**

Explicación práctica do análise mediante R dos datos relativos ao uso de licencias de Microsoft Office na Xunta de Galicia.

Partimos de varias fontes de datos (inventario de software dos equipos, correspondencias entre equipos e usuarios, información de características dos equipos segundo o Directorio Activo) e xeramos informes de análise de situación actual do uso de licencias de Microsoft Office que permiten verificar que non se superan o número de licencias permitidas e optimizar a súa distribución.

**Palabras e frases chave:** business reports, merge data, CMDB, ITSM

## 1. INTRODUCIÓN

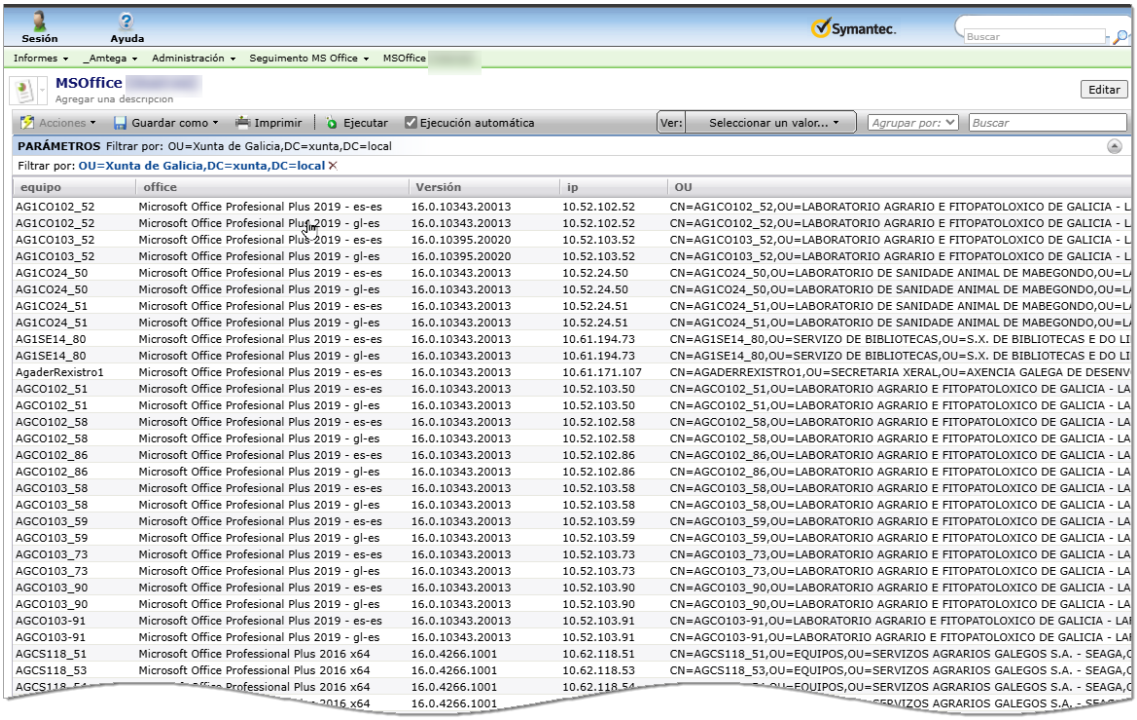
Partimos de varias fontes de datos:

- inventario de software dos equipos
- correspondencias entre equipos e usuarios
- información de características dos equipos segundo o Directorio Activo

Xeramos varios informes de análise de situación actual do uso de licencias de Microsoft Office.
Permiten verificar que non se superan o número de licencias permitidas.
E serven para optimizar a distribución interna (asignación) das licencias dentro das distintas unidades da organización: segundo consellerías, direccións xerais...

De esta maneira optimízase o uso dunhas aproximadamente 5.000 licencias sobre un conxunto potencial de 20.000 equipos (entre equipos de sobremesa e portátiles; con sistema operativo Windows).

## 2. Procesamento dos datos de inventario de software dos equipos




| equipo | office | Versión | ip | OU |
|---|---|---|---|---|
| AG1CO102_52 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.102.52 | CN=AG1CO102_52,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - L |
| AG1CO102_52 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.102.52 | CN=AG1CO102_52,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - L |
| AG1CO103_52 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10395.20020 | 10.52.103.52 | CN=AG1CO103_52,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - L |
| AG1CO103_52 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10395.20020 | 10.52.103.52 | CN=AG1CO103_52,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - L |
| AG1CO24_50 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.24.50 | CN=AG1CO24_50,OU=LABORATORIO DE SANIDADE ANIMAL DE MABEGONDO,OU=LA |
| AG1CO24_50 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.24.50 | CN=AG1CO24_50,OU=LABORATORIO DE SANIDADE ANIMAL DE MABEGONDO,OU=LA |
| AG1CO24_51 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.24.51 | CN=AG1CO24_51,OU=LABORATORIO DE SANIDADE ANIMAL DE MABEGONDO,OU=LA |
| AG1CO24_51 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.24.51 | CN=AG1CO24_51,OU=LABORATORIO DE SANIDADE ANIMAL DE MABEGONDO,OU=LA |
| AG1SE14_80 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.61.194.73 | CN=AG1SE14_80,OU=SERVIZO DE BIBLIOTECAS,OU=S.X. DE BIBLIOTECAS E DO LI |
| AG1SE14_80 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.61.194.73 | CN=AG1SE14_80,OU=SERVIZO DE BIBLIOTECAS,OU=S.X. DE BIBLIOTECAS E DO LI |
| AgaderRexistro1 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.61.171.107 | CN=AGADERREXISTRO1,OU=SECRETARIA XERAL,OU=AXENCIA GALEGA DE DESENV |
| AGCO102_51 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.50 | CN=AGCO102_51,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO102_51 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.50 | CN=AGCO102_51,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO102_58 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.102.58 | CN=AGCO102_58,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO102_58 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.102.58 | CN=AGCO102_58,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO102_86 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.102.86 | CN=AGCO102_86,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO102_86 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.102.86 | CN=AGCO102_86,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_58 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.58 | CN=AGCO103_58,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_58 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.58 | CN=AGCO103_58,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_59 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.59 | CN=AGCO103_59,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_59 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.59 | CN=AGCO103_59,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_73 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.73 | CN=AGCO103_73,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_73 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.73 | CN=AGCO103_73,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_90 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.90 | CN=AGCO103_90,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103_90 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.90 | CN=AGCO103_90,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103-91 | Microsoft Office Profesional Plus 2019 - es-es | 16.0.10343.20013 | 10.52.103.91 | CN=AGCO103-91,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCO103-91 | Microsoft Office Profesional Plus 2019 - gl-es | 16.0.10343.20013 | 10.52.103.91 | CN=AGCO103-91,OU=LABORATORIO AGRARIO E FITOPATOLOXICO DE GALICIA - LA |
| AGCS118_51 | Microsoft Office Professional Plus 2016 x64 | 16.0.4266.1001 | 10.62.118.51 | CN=AGCS118_51,OU=EQUIPOS,OU=SERVIZOS AGRARIOS GALEGOS S.A. - SEAGA,O |
| AGCS118_53 | Microsoft Office Professional Plus 2016 x64 | 16.0.4266.1001 | 10.62.118.53 | CN=AGCS118_53,OU=EQUIPOS,OU=SERVIZOS AGRARIOS GALEGOS S.A. - SEAGA,O |
| AGCS118_54 | Microsoft Office Professional Plus 2016 x64 | 16.0.4266.1001 | 10.62.118.54 | CN=AGCS118_54,OU=EQUIPOS,OU=SERVIZOS AGRARIOS GALEGOS S.A. - SEAGA,O |
| | 2016 x64 | 16.0.4266.1001 | | SERVIZOS AGRARIOS GALEGOS S.A. - SEAGA,O |

```r
logger::log_info("leyendo fichero {fname_instalaciones_office}...")
instalado_0 <- read.table(
  fname_instalaciones_office,
  sep = ",",
  header = TRUE,
  fileEncoding = "UTF-8-BOM"
) %>%
  as_tibble() %>%
  rename_with(~ if_else(. %in% c("Versión", "Version"), "version", .)) %>%
  verify(nrow(.) %in% 6000:8000) %>%
  filter(
    !str_detect(
      office,
      "zuzenketa|Add-in|ScreenTip|Activation|Language Pack|Converter|SharePoi
    )
  ) %>%
  mutate(office = str_remove(office, "Microsoft Office "))
instalado_0

equiv_office_versions <- tribble( ~nversion, ~txt_version,
                                  9,   "Office 2000",
                                  10,   "Office XP",
                                  11,   "Office 2003",
                                  12,   "Office 2007",
                                  14,   "Office 2010",
                                  15,   "Office 2013",
                                  16,   "Office 2016/2019/2021"
)
instalado <- instalado_0 %>%
  select(-OU) %>%
  mutate(
    nversion = str_replace(version, "^(\\d{1,3}).*$", "\\1") %>%
      as.numeric(),
    tiene_office = !is.na(version)
  ) %>%
  #mutate(OU8 = OU8_corregido) %>%
  left_join(
    equiv_office_versions,
    by(nversion)
```

## 3. Correspondencias entre equipos e usuarios

| Nombre | Dirección IP | Usuario | Name1 | OUAD |
|---|---|---|---|---|
| EQUIP043254 | 10.50.14 | cpazpan | Pazo Paniagua, | CN=EQUIP043254,OU=SERVIZO DE CONCILIAC |
| EQUIP043255 | 10.50.20 | mbprogon | Prol Gonzalez, | CN=EQUIP043255,OU=SERVIZO DE CONCILIAC |
| EQUIP043256 | 10.50.14 | myebbiu | Yebra Biurrun, | CN=EQUIP043256,OU=SERVIZO DE CONCILIAC |
| EQUIP043257 | 10.50.14 | icrugar | Cruz García, Is | CN=EQUIP043257,OU=XERENCIA,OU=AXENCIA |
| EQUIP043258 | 10.50.70 | jbardia | Barreiro Díaz, | CN=EQUIP043258,OU=D.X. DE SIMPLIFICACI |
| EQUIP043259 | 10.50.14 | cmcalrus | Calviño Ruso, C | CN=EQUIP043259,OU=SERVIZO DE CONCILIAC |
| EQUIP043260 | 10.50.20 | jabatiz | Abad Tizón, Jos | CN=EQUIP043260,OU=SERVIZO DE APOIO A F |
| EQUIP043261 | 10.50.41 | mdfervaz | Ferreiro Vázque | CN=EQUIP043261,OU=SERVIZO DE VALORACIO |
| EQUIP043262 | 10.50.41 | miglarm2 | Iglesias Armada | CN=EQUIP043262,OU=SERVIZO DE OBRAS,OU= |
| EQUIP043263 | 10.50.14 | mpsangon | Sanjurjo Gonzál | CN=EQUIP043263,OU=SERVIZO DE CONCILIAC |
| EQUIP043264 | 10.50.14 | alamces | Lameiro Ces, An | CN=EQUIP043264,OU=S.X. DE COORDINACION |
| EQUIP043265 | 10.50.20 | igonigl | González Iglesi | CN=EQUIP043265,OU=SERVIZO DE APOIO A F |
| EQUIP043266 | 10.50.14 | epertor | Peréz Torre, Ev | CN=EQUIP043266,OU=SERVIZO DE APOIO A F |
| EQUIP043267 | 10.50.14 | mmougon | Moure González, | CN=EQUIP043267,OU=SERVIZO DE CONCILIAC |
| EQUIP043269 | 10.50.20 | jcamvil | Cameselle Villa | CN=EQUIP043269,OU=SERVIZO DE CONCILIAC |
| EQUIP043270 | 10.50.20 | avazsan1 | Vazquez Sanmart | CN=EQUIP043270,OU=S.X. DE DEMOGRAFIA E |
| EQUIP043271 | 10.50.14 | frodviv | Rodríguez Vivia | CN=EQUIP043271,OU=SERVIZO DE PLANIFICA |
| EQUIP043272 | 10.50.13 | mrojote | Rojo Otero, Mar | CN=EQUIP043272,OU=GABINETE,OU=CONSELLE |
| EQUIP043273 | 10.50.13 | mdocper2 | Docampo Pérez, | CN=EQUIP043273,OU=GABINETE,OU=CONSELLE |
| EQUIP043276 | 10.50.16 | erioban | Del Río Baña, E | CN=EQUIP043276,OU=GABINETE,OU=CONSELLE |
| EQUIP043277 | 10.50.13 | amanper | Manzano Pérez, | CN=EQUIP043277,OU=CONSELLEIRO,OU=CONSE |
| EQUIP043278 | 10.50.13 | fmarrod | Márquez Rodrígu | CN=EQUIP043278,OU=CONSELLEIRO,OU=CONSE |
| EQUIP043279 | 10.50.13 | mgomrod1 | Gómez Rodriguez | CN=EQUIP043279,OU=CONSELLEIRO,OU=CONSE |
| EQUIP043280 | 13 | amirram | Miranda Ra | 043280,OU=GABINETE,OU=CONSELLE |
| | | cpitmar | Bi | =SERVIZO DE XESTI |

## 3. Información de características dos equipos segundo o Directorio Activo

| Name | IPv4Address | DistinguishedName | D | OperatingSystem | OperatingSystemVersion | La |
|---|---|---|---|---|---|---|
| EQUIP121468 | 10.54.142. | CN=EQUIP121468,OU=Museo Arqueoloxic | u | Windows 10 Pro | 10.0 (19045) | 05 |
| EQUIP043084 | 10.51.16 | CN=EQUIP043084,OU=Departamento de F | x | Windows 11 Pro | 10.0 (22631) | 21 |
| EQUIP020604 | 10.51.26 | CN=EQUIP020604,OU=Departamento de M | i | Windows 11 Pro | 10.0 (22621) | 02 |
| EQUIP161619 | 10.56.11 | CN=EQUIP161619,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 19 |
| EQUIP161618 | 10.56.11 | CN=EQUIP161618,OU=Equipos Xustiza,O | | Windows 11 Pro | 10.0 (22631) | 18 |
| EQUIP161617 | 10.64.225 | CN=EQUIP161617,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 21 |
| EQUIP161621 | 10.62.81 | CN=EQUIP161621,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 19 |
| EQUIP161620 | 10.62.84 | CN=EQUIP161620,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 15 |
| EQUIP161622 | 10.53.23 | CN=EQUIP161622,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 15 |
| EQUIP161623 | 10.52.14 | CN=EQUIP161623,OU=Equipos Xustiza,O | u | Windows 11 Pro | 10.0 (22631) | 19 |
| EQUIP079206 | 10.57.11 | CN=EQUIP079206,OU=Escola Oficial Na | a | Windows 11 Pro | 10.0 (22631) | 21 |
| EQUIP050110 | 10.50.36 | CN=EQUIP050110,OU=Servizo de Obra d | j | Windows 10 Pro | 10.0 (19045) | 20 |
| EQUIP121462 | 10.53.14 | CN=EQUIP121462,OU=Museo do Castro d | r | Windows 11 Pro | 10.0 (22631) | 20 |
| EQUIP121442 | 10.53.14 | CN=EQUIP121442,OU=Museo do Castro d | a | Windows 11 Pro | 10.0 (22631) | 18 |
| EQUIP121444 | 10.53.14 | CN=EQUIP121444,OU=Museo do Castro d | m | Windows 11 Pro | 10.0 (22631) | 21 |
| EQUIP121445 | 10.53.14 | CN=EQUIP121445,OU=Museo do Castro d | d | Windows 11 Pro | 10.0 (22631) | 17 |
| EQUIP121453 | | CN=EQUIP121453,OU=Arquivo do Reino | n | Windows 11 Pro | 10.0 (22631) | 22 |
| EQUIP121454 | | CN=EQUIP121454,OU=Arquivo do Reino | s | Windows 11 Pro | 10.0 (22631) | 21 |
| EQUIP121456 | 10.62.30 | CN=EQUIP121456,OU=Biblioteca Public | m | Windows 11 Pro | 10.0 (22631) | 20 |
| EQUIP121458 | 10.62.30 | CN=EQUIP121458,OU=Biblioteca Public | m | Windows 11 Pro | 10.0 (22631) | 19 |
| EQUIP080432 | 10.53.16 | CN=EQUIP080432,OU=Servizo de Gardac | r | Windows 10 Pro | 10.0 (19045) | 21 |
| EQUIP135357 | 192.168.( | CN=EQUIP135357,OU=Servizos Centrais | t | Windows 10 Pro | 10.0 (19045) | 06 |
| EQUIP137856 | 10.64.10 | CN=EQUIP137856,OU=Ourense,OU=Equipo | E | Windows 10 Pro | 10.0 (19045) | 21 |
| EQUIP135128 | | CN=EQUIP135128,OU=Equipos Xustiza,O | u | Windows 10 Pro | 10.0 (19045) | 01 |
| EQUIP137232 | | CN=EQUIP137232,OU=Equipos Xustiza,O | u | Windows 10 Pro | 10.0 (19045) | 29 |
| EQUIP031053 | 10.61.34 | CN=EQUIP031053,OU=D.X. de Xuventude | l | Windows 11 Pro | 10.0 (26100) | 21 |
| EQUIP | | CN=EQUIP010639,OU=Departamento de | | | 10.0 (19045) | 12 |
| | | EQUIP031051,OU=D.X. de | | | 10.0 (26100) | |

## 4. Cruce de datos entre distintos orígenes

```r
# cruce entre equipos según AD, instalaciones Office y Symantec Agent ====
cruce <- equipos_AD %>%
  #select(equipo, usuario, usuario_generico) %>%
  full_join(
    instalado %>%
      select(equipo, ip, tiene_office, office, txt_version) %>%
      group_by(equipo) %>%
      summarise(
        office = str_c(unique(office), collapse = ", "),
        tiene_office = any(tiene_office),
        txt_version = str_c(unique(txt_version), collapse = ", "),
        ip = str_c(unique(ip), collapse = ", ")
      ) %>%
      ungroup(),
    by = "equipo"
  ) %>%
  replace_na(list(tiene_office = FALSE)) %>%
  left_join(
    SA_equipos %>%
      group_by(equipo) %>%
      filter(row_number() == 1) %>%
      ungroup() %>%
      select(equipo, usuario_SA = usuario, usuario_generico_SA),
    by = "equipo"
  ) %>%
  mutate(tiene_symantec_agent = !is.na(usuario_generico_SA)) %>%
  verify(!duplicated(equipo))

cruce %>%
  count(tiene_office, usuario_generico)
```

```r
# recuento equipos según tengan Symantec Agent ====
cruce %>%
  replace_na(list(OU7 = "-")) %>%
  count(OU7, tiene_symantec_agent) %>%
  mutate(tiene_symantec_agent = if_else(tiene_symantec_agent, "con_SA", "sin_SA")) %>%
  pivot_wider(names_from = "tiene_symantec_agent", values_from = "n") %>%
  janitor::adorn_totals("row") %>%
  mutate(pct_sin_SA = formattable::percent(sin_SA / (sin_SA + con_SA), digits = 0)) %>%
  arrange(OU7 == "Total", desc(pct_sin_SA))

situacion_resumen <- cruce %>%
  count(OU7, OU6, tiene_office, usuario_generico, tiene_symantec_agent) %>%
  filter(!usuario_generico & tiene_symantec_agent) %>%
  verify(!usuario_generico) %>%
  group_by(OU7, OU6) %>%
  summarise(
    con_office = sum(if_else(tiene_office & !usuario_generico, n, 0L)),
    equipos = sum(n),
    .groups = "drop"
  ) %>%
  verify(!is.na(OU7))
situacion_resumen %>%
  janitor::adorn_totals("row") %>%
  print(n = Inf)
```

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# APRENDENDO ESTATÍSTICA CON LEARNINGSTATS

Elena Fernández-Sedes, María Isabel Borrajo[1,2] e Mercedes Conde-Amboage[1,2]

[1] Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga), 15705 Santiago de Compostela, España.
[2] Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela, España.

## RESUMO

Nos últimos anos incrementouse notablemente o interese polo coñecemento e a análise de datos en ámbitos moi diversos como a Economía, as Ciencias da saúde ou o Deporte. Este feito converteu o coñecemento e o manexo da Estatística nunha competencia imprescindible para as/os profesionais destes ámbitos. Co obxectivo de facilitar o proceso de ensino-aprendizaxe da Estatística desenvolveuse en R o paquete `LearningStats`, que se presenta como unha ferramenta didáctica versátil e rigorosa, que busca fomentar a comprensión e interpretación dos procedementos estatísticos sen importar o ámbito de desenvolvemento dos mesmos. A súa finalidade didáctica fai que se priorize a intuitividade e a sinxeleza de uso, ofrecendo tamén representacións gráficas e mensaxes de erro ou aviso facilmente comprensibles.

O paquete `LearningStats` inclúe as funcións necesarias para traballar nas diferentes áreas da Estatística: Estatística Descritiva, Modelos de Probabilidade, Inferencia Estatística e Regresión. Ademais, incorpora conxuntos de datos que permiten ilustrar o uso na práctica dos distintos procedementos e ferramentas.

**Palabras e frases chave**: Estatística, paquete de R, programación.

## 1. INTRODUCIÓN

Este traballo centrarase na presentación e exemplificación de tres funcións do paquete `LearningStats`: dúas destinadas a contrastar a bondade de axuste de modelos de distribución, e outra función que ofrece representacións gráficas para comparar visualmente modelos de distribución teóricos xunto coa distribución estimada a partir dunha certa mostra, así como as funcións de densidade ou masa de probabilidade.

## 2. CONTRASTES DE BONDADE DE AXUSTE

Os contrastes de bondade de axuste sobre modelos de distribución son procedementos que permiten avaliar se os datos dunha certa mostra se axustan a certa distribución de probabilidade. Os métodos de contraste a considerar neste proxecto son o test de Kolmogorov-Smirnov, o test de Lilliefors e o test chi-cadrado, todos vistos na materia de *Inferencia Estatística* que se imparte no terceiro curso do Grao en Matemáticas da Universidade de Santiago de Compostela.

Estes métodos implementáronse en dúas funcións que se poden executar mediante os comandos `gof.ks.test` e `gof.chisq.test`. O primeiro deles realiza o test de Kolmogorov-Smirnov (KS) mentres que o segundo leva a cabo un test chi-cadrado. Se ben o primeiro só é válido cando se consideran distribucións de probabilidade continuas, o segundo permite traballar con calquera tipo de distribución, sexa continua ou discreta. Ademais ambas funcións contemplan o caso no que a

distribución non estea completamente especificada, obrigando así a que se estimen os parámetros, ben a través do método dos momentos ou por máxima verosimilitude. É precisamente nestes casos onde a función `gof.ks.test` estende a idea do test de Lilliefors a outras distribucións.

Proporcionados os argumentos de entrada, ambas funcións proceden de forma similar: calculan o valor observado do estatístico de contraste xunto co p-valor ou nivel crítico asociado e proporcionan unha representación gráfica da distribución do estatístico baixo a hipótese nula, amosando o valor observado na mostra e a rexión de rexeitamento. Isto permite que a/o usuaria/o avalíe graficamente se existen evidencias estatisticamente significativas para rexeitar a hipótese nula.

En canto aos estatísticos de contraste, cada test mencionado presenta o seu. No caso do test KS, o estatístico recibe o nome de estatístico de Kolmogorov-Smirnov. Este é de distribución libre e, denotando por $F_n$ a función de distribución empírica da mostra, a súa expresión é

$$D_n = \underset{x\in\mathbb{R}}{sup}\,|F_n(x) - F_\theta(x)|.$$

Cando se estiman os parámetros, é necesario substituír na expresión anterior o parámetro $\theta$ pola estimación correspondente, $\widehat{\theta}$, de xeito que o estatístico deixa de ter distribución libre.

No test chi-cadrado, o estatístico calcúlase empregando a diferenza de frecuencias observadas e esperadas en cada categoría. Así, se a mostra se divide en $k$ categorías e se denotan por $O_i$ e $E_i$ as frecuencias observada e esperada en cada unha delas respectivamente, o estatístico vén dado por:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Como se ve en [2], o estatístico $T$ segue unha distribución chi-cadrado cuxos graos de liberdade dependen tanto do número de categorías como da estimación de parámetros.

A continuación proporciónanse dous exemplos.

**Exemplo 1.** Implementación do test de Kolmogorov-Smirnov con parámetros coñecidos.

Executando o código:

```
datos <- c(3.8, 0.1, 1, 0.25, 1.8, 3.9, 0.2, 3.5, 3, 0.3)
gof.ks.test(x = datos, dist = 'unif', theta = c(0,4))
```

obtense a saída seguinte e a representación gráfica que se amosa na Figura 1,

```
  Kolmogorov-Smirnov test (known parameters)

H0: data come from a Unif(0,4) distribution
Ha: data do not come from a Unif(0,4) distribution

data: datos
Dn = sup|Fn(x)-F0(x)|
alpha = 0.05
RR = (0.412, +Inf)
d_n,obs = 0.325, n = 10, p-value = 0.1885
```

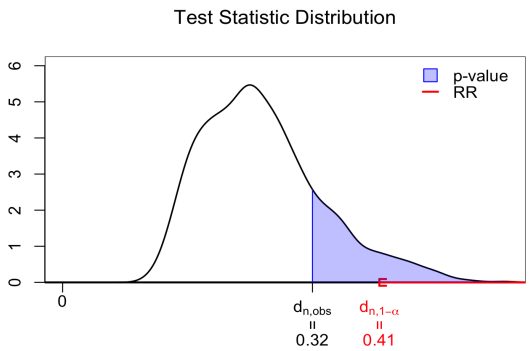



Figura 1: distribución de $D_n$, rexión crítica e p-valor asociados ao Exemplo 1.

**Exemplo 2.** Implementación do test chi-cadrado con parámetros descoñecidos.

Executando o código:

```
datos <- c(22, 53, 58, 39, 20, 5, 2, 1)
gof.chisq.test(x = datos, dist = "pois", xnames = 0:7)
```

obtéñense a saída que se mostra a continuación e a representación gráfica que se presenta na Figura 2,

```
 Chi-squared test for goodness-of-fit

H0: data follows 'pois' distribution
Ha: data does not follow 'pois' distribution

Estimated parameters:
[1] 2.05

Chi.squared = sum {(Oi - Ei)^2 / Ei}
Chi.squared follows X^2_{k-1-q}
alpha = 0.05
t_obs = 2.03967
RR = (9.48773, +Inf)
p-value = 0.72846

Observed frequencies, Oi:
{0}        {1}        {2}        {3}        {4}   {5,6,7,...}
22         53         58         39         20             8

Expected frequencies, Ei:
{0}        {1}        {2}        {3}        {4}   {5,6,7,...}
25.75      52.78      54.10      36.97      18.95        11.46

Warning message:
In gof.test(x = datos, dist = "pois", xnames = 0:7) :
Grouping has been carried out as some Ei was less than 5
```

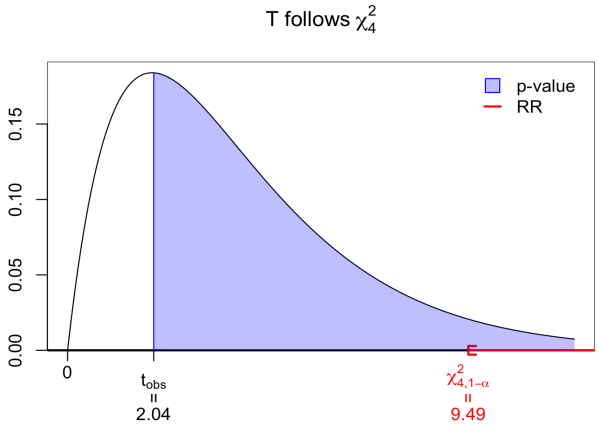



Figura 2: distribución de $T$, rexión de rexeitamento e p-valor para os datos do Exemplo 2.

## 3. COMPARACIÓN GRÁFICA ENTRE A ESTIMACIÓN DA FUNCIÓN DE DISTRIBUCIÓN E POSIBLES DISTRIBUCIÓNS POBOACIONAIS.

A función `plotDistribution` permite comparar visualmente como se axustan certos datos numéricos a unha distribución específica. En concreto pode representar as funcións de distribución, tanto a teórica indicada pola/o usuaria/o, coma a correspondente función estimada a partir da mostra proporcionada, e as respectivas funcións de densidade (ou masa de probabilidade se a distribución é discreta), permitindo así que a/o usuaria/o teña unha certa intuición antes de realizar un contraste de bondade de axuste como os que se presentan na Sección 2.

Para realizar ditas representacións gráficas empréganse ferramentas como o histograma, a función

de distribución empírica ou estimadores tipo núcleo da densidade. Ademais a función incorpora as estimacións dos parámetros que resulten precisos en cada caso. Preséntase a continuación un exemplo coa distribución normal:

**Exemplo 3.** Ilustración do comportamento da función `plotDistribution`.

Mediante o código,

```
set.seed(123)
sample.norm = rnorm(25, mean = 0, sd = 2)
plotDistribution(sample.norm, "norm", theta = c(0,2))
```

obtéñense as representacións gráficas da Figura 3.

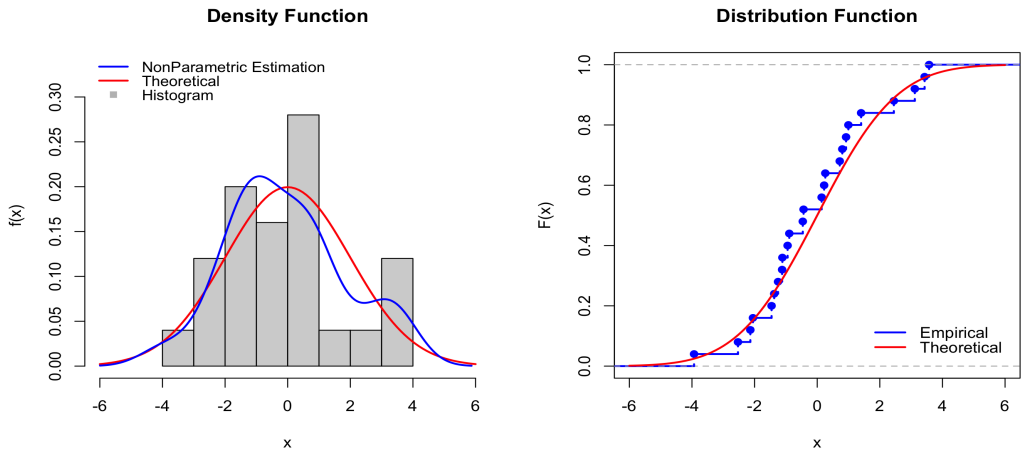



Figura 3: representacións gráficas obtidas mediante a función `plotDistribution` para os datos do Exemplo 3.

# Referencias


[1] Borrajo, M. I., Conde-Amboage, M. e López-Pérez, A. (2025). LearningStats: Elemental Descriptive and Inferential Statistics. R package version 0.1.0, `https://CRAN.R-project.org/package=LearningStats`

[2] A. García Pérez e Vélez Ibarrola. (1993). *Principios de Inferencia Estadística*. Universidad Nacional de Educación a Distancia (UNED).

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# MODELAGEM PARA OTIMIZAÇÃO DE ESCALAS DE FISCAIS DE PROVA

Luciane Ferreira Alcoforado[1] e João Paulo Martins dos Santos[2]

[1,2]Academia da Força Aérea Brasileira

## RESUMO

A elaboração das escalas de fiscalização das provas finais na Divisão de Ensino da Academia da Força Aérea (AFA) envolve desafios específicos, como a exigência de que o docente responsável pela disciplina não participe da aplicação da respectiva avaliação. Além disso, essas escalas são formuladas semanalmente, respeitando restrições como disponibilidade dos docentes, carga horária, distribuição equitativa de tarefas e horários das provas. Para enfrentar essa complexidade, foi desenvolvido um modelo de otimização linear inteira, capaz de gerar escalas mais eficientes e justas. A alimentação do modelo foi viabilizada por meio da automação do processo de organização dos dados, utilizando *scripts* em linguagem R. Essa automação gerava, para cada prova, uma lista de docentes elegíveis com base nas restrições estabelecidas, permitindo ao escalante tomar a decisão final de forma manual. A modelagem proposta, por sua vez, substitui esse processo por uma abordagem sistemática e imparcial, eliminando subjetividades e otimizando a alocação dos fiscais. Até o momento, os testes realizados foram conduzidos com base em dados fictícios, com o objetivo de verificar se o modelo era capaz de entregar soluções compatíveis com os critérios esperados. Embora os resultados iniciais tenham sido promissores, ainda não foram aplicados em situações reais, sendo necessária uma etapa posterior de validação prática para confirmar sua eficácia operacional.

**Palabras e frases chave**: Otimização linear inteira, Escalas de fiscalização, Alocação de docentes, Modelagem matemática.

## 1. INTRODUÇÃO

O problema abordado neste trabalho emerge diretamente da rotina acadêmica da Divisão de Ensino da Academia da Força Aérea (AFA), responsável pela condução de três Cursos de Formação de Oficiais [1]: CFOAV (Aviadores); CFOINF (Infantaria) e CFOINT (Intendentes). O corpo discente é distribuído em quatro Esquadrões, correspondentes aos anos de formação, cuja programação acadêmica exige sincronização entre atividades curriculares e extracurriculares.
Dentre as atividades regulares, destaca-se a aplicação de avaliações escritas formais, que requerem a presença do fiscal de provas (docente ou instrutor que não ministra a disciplina em questão). A construção dessas escalas de fiscalização é realizada semanalmente e deve respeitar múltiplas restrições, como disponibilidade individual, carga horária, alocação em aula e indisponibilidades previamente declaradas. O conjunto de dados utilizado para representar esse contexto operacional é composto por planilhas no formato *.xlsx*, incluindo: (i) quatro arquivos com as programações semanais de cada Esquadrão; (ii) uma planilha com os docentes e instrutores disponíveis para fiscalização; (iii) uma planilha com a distribuição das disciplinas ministradas por instrutor, turma e esquadrão; e (iv) uma planilha de indisponibilidades por dia, preenchida pelos próprios docentes. Historicamente, o processo de construção da escala era realizado manualmente, exigindo inspeção individual de restrições para cada potencial fiscal. A automação inicial, implementada via *scripts*

[1]Consulte o Link para mais informações.

em linguagem R, permitia consolidar os dados em um único *dataframe* e gerar, para cada avaliação, uma lista de candidatos elegíveis com base nas restrições operacionais. A decisão final, no entanto, permanecia sob responsabilidade do gestor, o que introduzia riscos de erro humano e viés subjetivo. Com o objetivo de aprimorar esse processo, propôs-se a formulação de um modelo de programação linear inteira, capaz de gerar automaticamente a escala de fiscalização, minimizando o custo total de alocação e assegurando imparcialidade na seleção dos fiscais. A modelagem foi validada em uma base de dados fictícia, permitindo verificar a aderência das soluções geradas aos critérios esperados, embora ainda não tenha sido aplicada em ambiente operacional real.

## 2. FORMULAÇÃO DO MODELO MATEMÁTICO

O problema de alocação é formalizado como um problema de PLI (Programação Linear Inteira), baseado nos autores [1] e [2].

### 2.1. Conjuntos

- $I$: Conjunto de docentes, indexado por $i$.
- $J$: Conjunto de dias da semana, indexado por $j$.
- $K$: Conjunto de horários de início de fiscalização ($K = \{7h, 8h, 13h\}$), indexado por $k$.
- $A$: Conjunto de provas a serem fiscalizadas, indexado por $a$.

### 2.2. Parâmetros

- $DFa$: Duração da fiscalização da prova $a$ em horas (1 ou 2 horas).
- $DA_a$: Disciplina associada à prova $a$.
- $DP_a$: O dia da semana específico ($j \in J$) em que a prova $a$ será realizada.
- $HIP_a$: O horário de início específico ($k \in K$) em que a prova $a$ será realizada.
- $d_{i,j,k}$: Variável binária. É 1 se o docente $i$ está disponível para iniciar uma fiscalização no dia $j$ e horário $k$; 0 caso contrário.
- $e_{i,a}$: Variável binária. É 1 se o docente $i$ leciona a 'disciplina' relacionada à prova $a$; 0 caso contrário.
- $CD_i$: Carga horária didática atual do docente $i$.
- $CF_i$: Carga horária total de fiscalizações que o docente $i$ já realizou em um período anterior.
- $w_F$: Peso atribuído à duração da fiscalização na função objetivo. Um valor usual é 1.0.
- $w_D$: Peso atribuído à carga horária didática do docente na função objetivo, para priorização.
- $w_A$: Peso atribuído ao histórico de fiscalizações anteriores do docente na função objetivo, para priorização.
- $fs$: Número inteiro, o máximo de fiscalizações que um docente pode realizar em uma única semana. Por exemplo, 1, 2, etc.

### 2.3. Variável de Decisão

Como o dia e o horário de cada prova são fixos, a variável de decisão é formulada como:
$X_{i,a}$: Variável binária. É 1 se o docente $i$ é alocado para fiscalizar a prova $a$; 0 caso contrário.

### 2.4. Função Objetivo

O objetivo é minimizar a soma ponderada dos custos de alocação, incorporando a carga horária da fiscalização e os fatores de priorização (carga didática e fiscalizações anteriores dos docentes):

$$\text{Min} \sum_{i \in I} \sum_{a \in A} (w_F \cdot DF_a + w_D \cdot CD_i + w_A \cdot CF_i) \cdot X_{i,a}$$

### 2.5. Restrições

1. **Restrição de Disponibilidade do Docente:** Um docente só pode ser alocado para uma prova se estiver disponível no horário e dia de início da prova.

$$X_{i,a} \leq d_{i,DP_a,HIP_a} \quad \forall i \in I, \forall a \in A$$

2. **Restrição de Não Fiscalizar Disciplina Própria:** Um docente não pode fiscalizar uma prova da disciplina que ele leciona.

$$X_{i,a} \leq 1 - e_{i,a} \quad \forall i \in I, \forall a \in A$$

3. **Restrição de Atribuição de Provas:** Cada prova deve ser fiscalizada por exatamente um docente.

$$\sum_{i \in I} X_{i,a} = 1 \quad \forall a \in A$$

4. **Restrição de Ocupação de Horários:** Um docente não pode ser alocado para mais de uma fiscalização simultaneamente. Para cada docente, em cada dia e em cada *slot* de tempo de 1 hora, a soma das alocações deve ser no máximo 1.

Considerando os horários de início fixos (7h, 8h, 13h) e as durações (1h ou 2h), cada fiscalização ocupa um ou dois blocos de 1 hora. Seja $T_{slots}$ o conjunto de todos os blocos de 1 hora no qual uma fiscalização pode ocorrer, ou seja, $T_{slots} = (\{7h, 8h, 9h, 13h, 14h, 15h\})$. Seja $S(a)$ o conjunto de todos os slots de tempo de 1 hora ocupados pela prova $a$, ou seja, $S(a)$ inclui o horário de início da prova $a$ ($HIP_a$) e, se a duração da prova for de 2 horas ($DF_a = 2$), também o slot de tempo subsequente ($HIP_a + 1$).

$$\sum_{a \in A | DP_a = j \, e \, t \in S(a)} X_{i,a} \leq 1, \quad \forall i \in I, \forall j \in J, \forall t \in T_{slots}$$

5. **Rotatividade Semanal:** Garante que cada docente seja alocado para no máximo um número $fs$ de fiscalizações por semana. Se o contexto exigir 'uma única vez de preferência', pode definir $fs = 1$.

$$\sum_{a \in A} X_{i,a} \leq fs \quad \forall i \in I$$

6. **Domínio das Variáveis:** As variáveis de decisão são binárias: $X_{i,a} \in \{0, 1\} \quad \forall i \in I, a \in A$.

## 3. IMPLEMENTAÇÃO COM R

### 3.1 Gerenciamento dos dados

Semanalmente há uma demanda por organizar a fiscalização de provas. O setor responsável elabora, para cada esquadrão, um quadro de horários (arquivo PDF) com todas as disciplinas e avaliações previstas para a semana. A partir desses documentos, são extraídas as informações relevantes como datas, horários e disciplinas para compor a lista de provas a serem aplicadas. A tarefa do escalante consiste em atribuir a cada prova um docente elegível, respeitando as restrições descritas. Para viabilizar a aplicação do modelo de otimização proposto, é necessário estruturar os dados em planilhas específicas, tais como: `docentes.csv`; `provas.csv`; `disponibilidade.csv` e `ensina_disciplina.csv`.

### 3.2 Implementação do modelo

O *script* a seguir implementa um Modelo de Programação Linear Inteira (PLI) em R, utilizando o pacote *lpSolve* e algumas funções de elaboração própria, para otimizar a alocação de docentes em fiscalizações de provas. Seu objetivo principal é minimizar a carga total de fiscalização, incorporando critérios de priorização para docentes com menor carga horária didática e menor histórico de fiscalizações anteriores, enquanto garante a conformidade com diversas restrições operacionais e regulamentares.

```
#1.1 Carregar e preparar dados
dados <- carregar_e_preparar_dados(path_docentes = "docentes.csv",
  path_provas = "provas.csv",path_disponibilidade =
  "disponibilidade.csv",  path_ensina = "ensina_disciplina.csv", path_pesos = "pesos.csv")
#1.2 Definir conjuntos e parâmetros a partir dos dados carregados
I <- dados$df_docentes$ID_Docente; A <- dados$df_provas$ID_Prova
J <- unique(dados$df_disponibilidade$Dia_Semana);
K <- unique(dados$df_disponibilidade$Horario_Inicio)
T_slots <- c("7h", "8h", "9h", "13h", "14h") # Horários de início
MaxFiscalizacoesSemanais <- 1 # Parâmetro de rotatividade
#1.3. Definir nomes e número de variáveis de decisão
var_names <- paste0("X_", dados$expanded_vars$ID_Docente, "_", dados$expanded_vars$ID_Prova)
num_vars <- length(var_names)
#2. Construir função objetivo
c_obj <- construir_funcao_objetivo(dados$expanded_vars, dados$pesos)
#3. Construir restrições
restricoes <- construir_restricoes(dados$expanded_vars, num_vars, I, A, J, T_slots,
    MaxFiscalizacoesSemanais)
#4. Definir tipos de variáveis e limites
types <- rep("binary", num_vars);
lower_bounds <- rep(0, num_vars); upper_bounds <- rep(1, num_vars)
#5. Resolver o Modelo
solucao <- lpSolve::lp(direction = "min", objective.in = c_obj,
  const.mat = restricoes$A_matrix, const.dir = restricoes$dir,
  const.rhs =  restricoes$rhs, int.vec = 1:num_vars,
  binary.vec = 1:num_vars, all.bin = TRUE)
#6. Analisar Resultados
analisar_resultados(solucao, var_names, dados$df_docentes, dados$df_provas)
```

### 3.3 Exemplo de Saída

A tabela 1 fornece os dados de entrada do modelo de otimização linear. Alocações Finais obtidas pelo modelo são fornecidas na tabela 2.

Táboa 1: Resumo dos Dados de Entrada dos Docentes: Carga horária (em horas), Fisc. Anteriores(em horas), Disponibilidade (Dia Hora).

| ID do Docente | Carga Didática | Fisc. Anteriores | Disciplinas que Leciona | Disponibilidade (Dia Hora) |
|---|---|---|---|---|
| D001 | 40 | 3 | Disciplina5, Disciplina6, Disciplina9 | Seg 13h |
| D002 | 24 | 5 | Disciplina2, Disciplina9, Disciplina10 | Sex 7h |
| D003 | 28 | 8 | Disciplina6, Disciplina10 | Seg 7h; Seg 13h; Qui 13h; Sex 7h; Sex 13h |
| D004 | 23 | 9 | Disciplina1, Disciplina4, Disciplina6 | Qui 7h; Qui 13h; Sex 7h; Sex 8h |
| D005 | 12 | 10 | Disciplina3, Disciplina6 | Seg 7h; Seg 8h; Seg 13h; Sex 8h |
| D006 | 19 | 4 | Disciplina3, Disciplina8 | Seg 13h; Qui 8h; Sex 8h; Sex 13h |
| D007 | 27 | 2 | Disciplina7 | Seg 7h; Qui 7h; Qui 13h; Sex 13h |
| D008 | 31 | 10 | Disciplina2, Disciplina7, Disciplina10 | Qui 7h; Sex 7h; Sex 13h |
| D009 | 20 | 8 | Disciplina5, Disciplina7 | Seg 7h; Seg 8h; Qui 8h; Sex 8h; Sex 13h |
| D010 | 14 | 8 | Disciplina5, Disciplina8 | Seg 13h; Qui 7h; Qui 13h |

Táboa 2: Alocação Otimizada de Docentes para Fiscalização de Provas.

| Docente | Prova | Disciplina | Duracao Horas | Dia | Horario Inicio | Horario Final |
|---|---|---|---|---|---|---|
| D003 | P001 | Disciplina9 | 2 | Seg | 7h | 9h |
| D005 | P004 | Disciplina10 | 1 | Seg | 7h | 8h |
| D009 | P002 | Disciplina3 | 1 | Qui | 8h | 9h |
| D006 | P005 | Disciplina7 | 2 | Qui | 8h | 10h |
| D007 | P003 | Disciplina8 | 2 | Sex | 13h | 15h |

## 4. CONCLUSÕES

Apresentamos neste trabalho uma aplicação prática do modelo de programação linear inteira aplicado ao problema de alocação de docentes às escalas semanais de fiscalização de provas. Mostramos como a utilização da ferramenta computacional R auxiliou em todo o processo, desde o recebimento dos arquivos em PDF, a extração e organização dos dados para alimentar os parâmetros do modelo até finalmente obter a solução ótima e retornar uma tabela com as informações organizadas para divulgação final aos docentes escalados. Esperamos que os resultados possam auxiliar na automação do processo de escalas.


## AGRADECIMENTOS

À Academia da Força Aérea e ao prof. Dr. Humberto José Lourenção pelas contribuições ao modelo.


# Referencias


[1] Arenales, M., Armentano, V., Morabito, R., Yanasse, H. (2011). *Pesquisa Operacional*. Elsevier : ABEPRO, Rio de Janeiro.

[2] Fávero, L.P., Belfiore, P. (2013). *Pesquisa Operacional para Cursos de Engenharia*. Elsevier, Rio de Janeiro.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

**Web scraping de páxinas dinámicas. Automatizando o buscador con RSelenium.**

Martín García Cebeiro[1]

[1]Universidade de Santiago de Compostela

**RESUMO**

A análise de datos é unha disciplina que move o mundo de hoxe, pero para poder desempeñala, primeiro é necesario obter ditos datos. Os problemas aparecen cando non se ten acceso directo a unha base de datos xa confeccionada. Afortunadamente, resulta común que os datos que se desexan analizar aparezan recollidos nalgunha páxina web. O termo *web scraping* úsase para referirse ao proceso de recoller, de forma automática, datos ou outros elementos dunha páxina web a partir do seu código html.
Na XI Xornada de Usuarios de R en Galicia, Álvaro Theotonio presentou no seu relatorio "R, Webscraping e Open Science. Esquivando balas en Matrix." o procedemento para facer web scraping de páxinas web estáticas en R co uso do paquete `rvest` [1], e un dos obstáculos que menciona é precisamente as páxinas web dinámicas. Estas páxinas usan JavaScript para xerar o seu contido de forma dinámica cando, por exemplo, se pica nun botón, sen realizar cambios na dirección url. Este traballo pretende esquivar esta bala axudándonos do paquete `RSelenium` [2], que permite interactuar co buscador automaticamente dende R.

**Palabras e frases chave**: páxina web estática, páxina web dinámica, web scraping, html, url, JavaScript.

# Referencias

[1] Wickham H (2025). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.5, https://rvest.tidyverse.org/.

[2] Harrison J, Kim J, Völkle J (2025). *RSelenium: R Bindings for 'Selenium WebDriver'*. R package version 1.7.9, https://docs.ropensci.org/RSelenium/.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# DA ABUNDANCIA Á EVIDENCIA: FLUXO DE TRABALLO PARA A ANÁLISE DIFERENCIAL DE DATOS ÓMICOS EN R

Julia García-Currás[1]

[1] Biostatech, Advice, Training & Innovation in Biostatistics, Ames; CITIC-Universidade da Coruña, A Coruña.

## RESUMO

As ómicas son disciplinas que analizan distintos tipos de moléculas dos organismos no seu conxunto, como son xenes, transcritos, proteínas, lípidos ou metabolitos, entre outros. Os datos ómicos son de alta dimensionalidade e un dos softwares máis utilizados para a súa análise é R. Tan estendido está o uso desta ferramente que dentro do ecosistema do propio R atópanse proxectos como *Bioconductor*, o cal proporciona paquetes reproducibles e específicos para cada tipo de ómica e análise, permitindo que estas sexan estandarizadas e transparentes. Unha das análises máis comúns con datos ómicos cuantitativos é a análise de abundancia diferencial, consistente na identificación das moléculas que presentan diferenzas entre as condicións ou grupos de estudo. Un fluxo de traballo típico para esta análise inclúe os seguintes pasos: preprocesado da matriz de cuantificación, identificación de moléculas diferencialmente abundantes con corrección por comparacións múltiples, caracterización funcional das anteriores e análise de agrupación. Este fluxo de traballo estándar permite identificar moléculas relevantes asociadas a condicións biolóxicas ou clínicas específicas, e o uso de R para a súa implementación axuda a acadar análises reproducibles e fiables.

**Palabras e frases chave:** ómicas, análise diferencial, bioinformática, *bioconductor*.

## 1. INTRODUCIÓN

Dende comezos deste século, a aparición, constante mellora e progresivo abaratamento das tecnoloxías de secuenciación e de análise masiva de moléculas biolóxicas deu lugar a aparición dunha nova disciplina en bioloxía: as **ómicas**, que estudan, de maneira global e masiva, diferentes capas moleculares dun organismo. Esta disciplina comprende tanto a xeración destes datos biolóxicos de grande dimensionalidade como a súa posterior análise estatística, a cal se pode facer grazas ao uso de diferentes ferramentas computacionais, entre as que destaca especialmente R.

As **vantaxes de R** como ferramenta de referencia na análise deste tipo de datos inclúen, entre outras, o feito de ser un software libre e gratuíto, cunha comunidade moi activa que desenvolve, mantén e mellora os diferentes paquetes de análise. Ademais, R conta con proxectos como *Bioconductor* que ofrece paquetes específicos e estandarizados para distintos tipos de datos ómicos, favorecendo a reproducibilidade. Cómpre salientar tamén a transparencia: as usuarias e usuarios desta linguaxe poden sempre consultar como se implementan os diferentes algoritmos, coñecer as opcións usadas e, en moitos casos, modificalas segundo as súas necesidades.

## 2. ANÁLISE DE ABUNDANCIA DIFERENCIAL

A análise de datos ómicos é diversa e depende en gran medida do tipo de moléculas analizadas e do deseño experimental. En xeral, para aquelas moléculas das que se pode cuantificar a súa abundancia, como é o ARN (transcriptómica), as proteínas (proteómica), os lípidos (lipidómica) ou os metabolitos (metabolómica), o procedemento máis común é a **análise de abundancia diferencial**, consistente en comparar para cada molécula analizada as súas cantidades entre dous ou máis grupos de estudo. Normalmente, compárase unha condición (como por exemplo, unha enfermidade) fronte a un control, co obxectivo de identificar aquelas moléculas alteradas en dita condición.

O fluxo de traballo adoita incluír pasos iniciais obrigatorios, como o **preprocesado dos datos**, e outros opcionais, como a **análise de agrupacións de moléculas ou mostras**, ou a avaliación das moléculas con abundancia diferencial en canto a súa **caracterización funcional** ou **capacidade de discriminación** das condicións de estudo. Ademais, o uso de R permite non só analizar os datos, senón tamén representar os resultados de forma clara e variada para facilitar a súa interpretación.

O fluxo de traballo estándar e os paquetes e/ou funcións de R asociadas a cada un destes pasos son os seguintes:

- **Preprocesado** dos datos para xestionar valores ausentes, filtrando por datos faltantes e imputando (`impute`[1]), e para eliminar variacións non atribuíbles á condición de estudo mediante a normalización (`normalyzerDE`[2], `vsn::justvsn`[3], `preprocessCore::normalize.quantiles`[4], `limma::normalizeCyclicLoess`[5]).
- **Identificación de moléculas diferencialmente abundantes** mediante probas estatísticas clásicas ou modelos lineais, empregando paquetes como `limma`[5] ou `DESeq2`[6] ou funcións básicas de R `stats::t.test`[7] e correxíndo os valores p resultantes por comparacións múltiples. Os resultados represéntanse mediante gráficos de volcán (`EnhanceVolcano`[8]).
- **Agrupacións** de mostras ou moléculas para descubrir subconxuntos con patróns similares, empregando técnicas non supervisadas coma a análise de compoñentes principais ou PCA (`pcaMethods::pca`[9], `factoextra`[10] para a representación gráfica dos resultados), ou o *clustering* xerárquico (`stats::dist`[7] e `stats::hclus`[7]). O resultado destas últimas represéntanse comunmente mediante mapas de calor e dendrogramas (`stats::heatmap`[7], `heatmaply`[11], `dendextend`[12], `circlize`[13]).
- **Análise funcional** ou de enriquecemento que permite determinar procesos ou funcións celulares enriquecidas entre as moléculas con abundancia diferencial (`clusterProfiler`[14], `AnnotationDbi`[15], `enrichplot`[16] para representar os resultados).
- Avaliación da **capacidade de discriminación** das moléculas con abundancia diferencial respecto aos grupos de estudo, usando modelos de aprendizaxe automática supervisados, incluídos en paquetes como `caret`[17], `glmnet`[18] ou `randomForest`[19].

## 3. CONCLUSIÓNS

Este fluxo de traballo estándar permite sacar o máximo proveito dos datos ómicos cuantitativos coa finalidade de identificar **biomarcadores candidatos** para unha determinada condición, que poden servir de diagnose, prognose ou diana farmacolóxica, así como identificar **funcións biolóxicas chaves** alteradas e implicadas no desenvolvemento dunha condición ou na resposta a unha determinada intervención. Ademais, este fluxo é **estensible e personalizable** para cada estudo e tipo de datos, podendo engadir, eliminar ou modificar metodoloxías segundo o deseño do mesmo.

Finalmente, **a utilización de R para analizar estes datos, e especialmente que estas análises se compartan** de forma pública en repositorios como GitHub, contribúe á **transparencia e reproducibilidade**, mellorando a comparabilidade e facilitando a integración e interpretación da información.


## AGRADECEMENTOS

Julia García-Currás é beneficiaria dunha beca predoctoral da Axencia Galega de Innovación, Xunta de Galicia, 2022-2026 (Ref. 23_IN606D_2022_2707220) e realiza un doutoramento industrial na compañía Biostatech e na Universidade da Coruña-CITIC. Moitas grazas a Gael Naveira Barbeito polo apoio e confianza.


**Referencias**


[1] Hastie T et al., (2024). *i*mpute: Imputation for microarray data. doi:10.18129/B9.bioc.impute

[2] Willforss J et al., (2018). NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression *Analysis Journal of Proteome Research.*

[3] Huber W et al., (2002). Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* Vol 18, S96-S104.

[4] Bolstad B (2024). preprocessCore: A collection of pre-processing functions. doi:10.18129/B9.bioc.preprocessCore

[5] Ritchie M et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* Vol 43(7), e47.

[6] Love M et al., (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* Vol 15: 550.

[7] R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[8] Blighe K et al., (2018). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

[9] Stacklies W et al., (2007). pcaMethods -- a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics* Vol 23, 1164-1167

[10] Kassambara A et al., (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

[11] Galili T et al., (2017). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics.*

[12] Galili T (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*.

[13] Gu Z (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*.

[14] Xu S et al., (2024). Using clusterProfiler to characterize multiomics data. *Nature Protocols*.

[15]Pagès H et al., (2025). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor.

[16] Yu G (2024). enrichplot: Visualization of Functional Enrichment Result.

[17] Kuhn M (2008). Building Predictive Models in R Using the caret Package. J*ournal of Statistical Software* Vol 28(5): 1–26.

[18] Friedman J et al., (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* Vol 33(1): 1-22.18] Liaw A et al., (2002). Classification and Regression by randomForest. *R News* Vol 2(3): 18-22.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# TRABALLAR CON DATOS DE XEITO LIMPO E EFICIENTE COA LIBRARÍA *tidyverse*

Mª José Ginzo Villamayor[1]

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

## RESUMO

O *tidyverse* é un conxunto de paquetes de R que comparte unha filosofía e unha estrutura de deseño común, pensada para facer o traballo con datos máis limpo, eficiente e reproducible. Baseado na noción de *tidy data* —cada columna é unha variable, cada fila unha observación—, o ecosistema do *tidyverse* integra ferramentas coherentes para a importación, transformación, visualización e comunicación dos datos.

Nesta presentación amosarase como empregar paquetes como `readr`, `dplyr`, `tidyr` e `ggplot2` para realizar un fluxo de traballo completo: desde a lectura de datos ata a súa visualización e resumo. Ilustrarase con exemplos prácticos e boas prácticas para mellorar a reproducibilidade e claridade das análises, facendo especial fincapé na integración co ecosistema R e na creación de código expresivo e transparente.

**Palabras e frases chave**: R, tidyverse, análise de datos, reproducibilidade, visualización, ciencia de datos.

## 1. INTRODUCIÓN

Nos últimos anos, o uso de R extendiuse de forma significativa nos ámbitos académico, científico e profesional, converténdose nunha ferramenta de referencia para a análise estatística, visualización e reproducibilidade. Dentro deste ecosistema, o *tidyverse*, e especialmente o paquete `dplyr`[1], destaca pola súa coherencia conceptual e pola capacidade para simplificar as tarefas comúns de manipulación de datos.

Un concepto clave no *tidyverse* é o dos **"tidy data"** (datos ordenados): cada columna representa unha variable, cada fila un caso ou observación, e cada celda contén un único valor. Este formato facilita a manipulación, agregación e visualización de datos, ademais de permitir un fluxo de traballo consistente con funcións de `dplyr` como `select()`, `filter()`, `mutate()` ou `group_by()`.

Un dos retos principais cando se traballa con datos reais é a súa heteroxeneidade: columnas con tipos incorrectos, valores ausentes, estruturas desordenadas ou datos replicados. O *tidyverse* aborda estes retos ofrecendo unha sintaxe consistente para transformacións comúns e fomentando un estilo de programación modular, transparente e facilmente documentable, favorecendo o traballo colaborativo e a reproducibilidade das análises.

Esta introdución proporciona o marco conceptual, e na sección seguinte descríbese de forma detallada o fluxo de traballo no *tidyverse*, mostrando como aplicar estas ferramentas a un conxunto de datos real de forma reproducible e eficiente.

[1]O paquete `dplyr` foi desenvolvido por Hadley Wickham de RStudio e é unha versión optimizada do seu paquete `plyr`. Non engade funcionalidades novas a R, xa que todo o que se pode facer con `dplyr` poderíase facer tamén coa sintaxe básica de R. A súa principal contribución é proporcionar unha "gramática"para a manipulación de `data frames` mediante verbos claros, permitindo comunicar de forma intuitiva o que se fai cos datos a outras persoas. Ademais, as funcións de `dplyr` son moi rápidas, implementadas en C++.

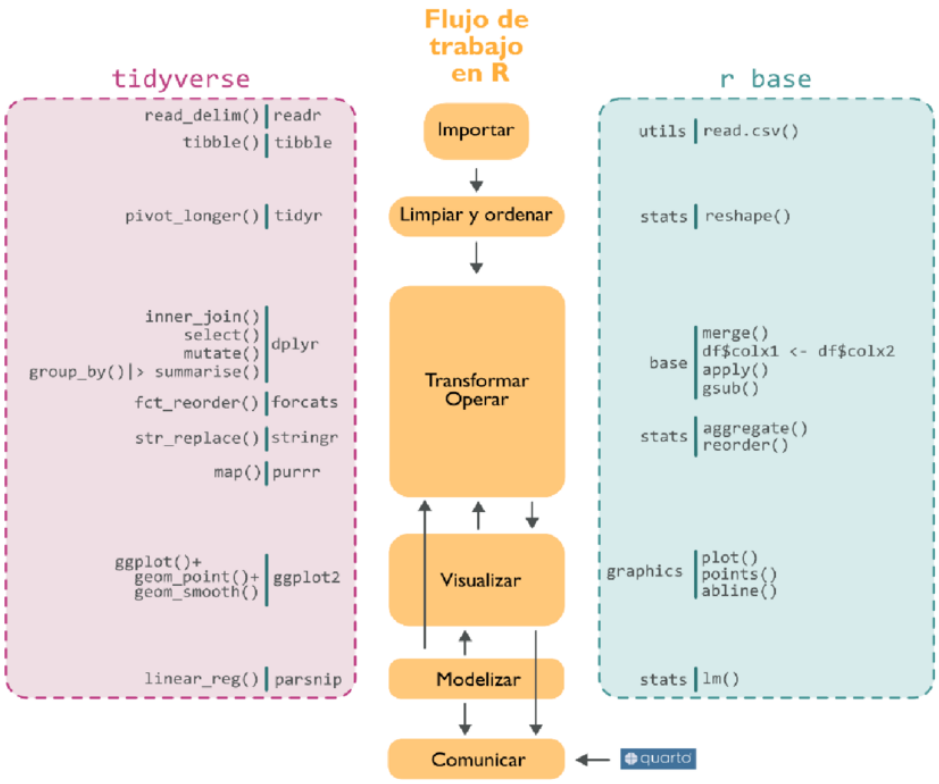



Figura 1: Fluxo de traballo no *tidyverse*: desde a lectura de datos, inspección, limpeza, transformación, agregación, visualización ata o reporte final. O diagrama compara tamén as funcións do *tidyverse* coas correspondentes de R base.

## 2. FLUXO DE TRABALLO E FERRAMENTAS PRINCIPAIS

O fluxo de traballo no *tidyverse* segue unha secuencia lóxica desde a lectura de datos, inspección, limpeza, transformación, agregación, visualización ata o reporte final (ver Figura 1). Este proceso faise principalmente usando funcións do paquete `dplyr`, que permiten seleccionar, filtrar, transformar, agrupar e resumir datos de forma eficiente.
Para ilustrar de forma máis sinxela cada etapa, na Figura 2 preséntase un fluxograma simplificado do proceso de manipulación de datos no *tidyverse*:
O fluxo de traballo pode describirse do seguinte xeito:

1. **Lectura:** con `read_csv()` ou `read_delim()` para importar os datos coas clases de columna adecuadas.
2. **Inspección inicial:** coas funcións `glimpse()`, `summary()` ou `head()` para coñecer a estrutura, valores extremos ou tipos de datos.
3. **Limpieza de datos faltantes:** usando `drop_na()`, `fill()` ou `replace_na()` para xestionar valores ausentes.
4. **Filtrado e selección:** `filter()` para quedarnos con filas relevantes, `select()` para eliminar columnas innecesarias.
5. **Creación de novas variables:** con `mutate()`, por exemplo para calcular medias, diferenzas ou transformadas de columnas existentes.
6. **Agrupación e resumo:** `group_by()` e `summarise()` para obter estatísticas agregadas segundo criterios xerais.

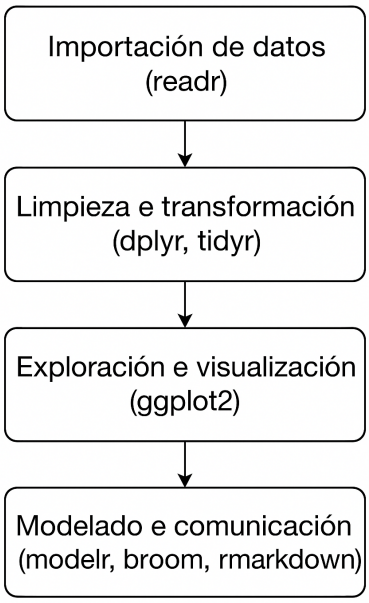



Figura 2: Fluxograma simplificado do proceso de manipulación de datos no *tidyverse*.

7. **Unión de datos:** con `left_join()` ou outras variantes para combinar con bases de datos externas ou adicionais.
8. **Reestruturación:** se é necesario, usar `pivot_longer()` ou `pivot_wider()` para adaptar o formato ao gráfico ou modelo.
9. **Uso do operador pipe (%¿%):** O operador pipe `%>%`, do paquete `magrittr`, permite encadear funcións de forma clara. O obxecto da esquerda pasa como primeiro argumento á función da dereita, mellorando a legibilidade do código. Exemplo:

   ```
   dados %>%
     select(idade) %>%
     filter(idade > 18) %>%
     summarise(media_idade = mean(idade, na.rm = TRUE))
   ```

   Se é necesario pasar o obxecto a un argumento diferente do primeiro, utilízase o punto (`.`) como marcador de posición.
10. **Visualización:** por exemplo:

    ```
    dados %>% group_by(categoria) %>%
      summarise(valor_medio = mean(valor, na.rm = TRUE)) %>%
      ggplot(aes(x = categoria, y = valor_medio)) +
      geom_col() +
      labs(title = "Exemplo de visualización agregada")
    ```

11. **Exportación ou reporte:** `write_csv()` ou creación dun informe reproducible con RMarkdown.

Tamén se poden incorporar figuras ou táboas que mostren os resultados antes e despois dunha transformación, ou comparativas entre métodos. Como se mostra na Figura 1, estas etapas seguen unha secuencia lóxica que facilita a reproducibilidade e a claridade na manipulación de datos.
Este fluxo será ilustrado na presentación mediante exemplos prácticos, mostrando paso a paso a aplicación das funcións do *tidyverse* sobre conxuntos de datos reais.

## 3. CONCLUSIÓNS

O *tidyverse* supón unha revolución conceptual na forma de traballar con datos en R. A súa coherencia interna, a súa sintaxe intuitiva e a súa ampla comunidade de usuarios contribúen a mellorar a reproducibilidade e a transparencia das análises. Ademais, a súa integración con outras ferramentas de R e co ecosistema de ciencia de datos en xeral convérteno nun recurso esencial tanto para a docencia como para a investigación aplicada.
O futuro do *tidyverse* pasa por seguir ampliando o seu alcance cara á modelización, o procesamento de datos espaciais e o traballo con grandes volumes de información, mantendo sempre o compromiso coa reproducibilidade e o código aberto.
En resumo, dominar o *tidyverse* non só facilita o traballo diario cos datos, senón que tamén prepara aos analistas e investigadores para afrontar retos crecentes en ciencia de datos de maneira eficiente, reproducible e colaborativa.


## AGRADECEMENTOS

Esta publicación forma parte da axuda PID2020-116587GBI00, financiada polo MCIN/AEI e polo FEDER, Unha maneira de facer Europa. Tamén contou co apoio da Xunta de Galicia, a través do Grupo de Referencia Competitiva ED431C 2021/24.


# Referencias


[1] Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. Dispoñible en: `https://r4ds.hadley.nz`

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

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

[4] Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4. Dispoñible en: `https://CRAN.R-project.org/package=dplyr`

[5] R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Viena, Austria. Dispoñible en: `https://www.R-project.org`

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Estimación de filamentos no plano mediante o uso da envoltura $r$-convexa. Aplicación a datos LiDAR

Héctor González-Vázquez[1], Beatriz Pateiro-López[1] e Alberto Rodríguez-Casal[1]

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

**RESUMO**

O problema de estimación de filamentos consiste en reconstruír ou estimar a estrutura de curva ou filamento arredor da cal se distribúen un conxunto de puntos. Neste traballo introducimos un estimador de filamentos que se basea no estimador EDT (*Euclidean Distance Transform*) proposto orixinalmente por [1] e o mellora empregando unha condición de forma coñecida como a $r$-convexidade, que xeneraliza a convexidade e proporciona maior flexibilidade ao estimador. En concreto, propoñemos empregar como estimador piloto do soporte dos datos a envoltura $r$-convexa da mostra [3]. Desenvolvemos unha implementación en R do estimador proposto facendo uso do paquete `alphahull` [2], que permite calcular a envoltura $r$-convexa dun conxunto finito de puntos no plano. Tamén aplicamos o noso estimador a uns datos LiDAR (*Light Detection And Ranging*) do ámbito forestal para estimar o contorno de seccións horizontais de troncos de árbores, para o que facemos uso do paquete `lidR` [4, 5].

**Palabras e frases chave**: Estimación de filamentos; Envoltura $r$-convexa; LiDAR; alphahull; lidR.

**AGRADECEMENTOS**

O primeiro autor agradece o apoio económico da axuda de apoio á etapa predoutoral da Xunta de Galicia ED481A-2025-173, cofinanciada parcialmente pola Unión Europea no marco do programa FSE+ Galicia 2021-2027. Este traballo é parte do proxecto de I+D+i PID2020-116587GB-I00 financiado por MICIU/AEI/10.13039/501100011033 e do proxecto ED431C 2025/03 financiado pola Consellería de Educación, Ciencia, Universidades e Formación Profesional.

## Referencias

[1] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L. (2012). The geometry of nonparametric filament estimation. *Journal of the American Statistical Association* 107(498), 788–799.

[2] Pateiro-López, B., Rodríguez-Casal, A. (2010). Generalizing the convex hull of a sample: the R package alphahull. *Journal of Statistical Software* 34(5), 1–28.

[3] Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques* 43(6), 763–774.

[4] Roussel J., Auty D. (2025). *Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. R package version 4.2.1, `https://cran.r-project.org/package=lidR`.

[5] Roussel J., Auty D., Coops N. C., Tompalski P., Goodbody T. R., Meador A. S., Bourdon J., de Boissieu F., Achim A. (2020). lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment* 251, 112061.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# SIMULACIÓN DE ESCENARIOS DE MERCADO MEDIANTE MODELADO ESTOCÁSTICO EN R

Daniel Grande Fernández[1], Antón Seijo Barrio[1]

[1] Universidade de Santiago de Compostela (USC)

## RESUMO

This paper presents an applied approach to market scenario simulation using the R programming language. Based on the theoretical foundation of stochastic modeling of asset prices, our emphasis is placed on the practical implementation of simulations and the insights they provide. We show how freely available financial data (in our case, from BBVA) can be downloaded and processed in R, and how simple statistical tools allow us to estimate the necessary parameters. We will generate a wide range of possible future market trajectories results are visualized through R's graphical libraries, which make the uncertainty of financial markets more tangible. The study highlights the value of computational experimentation: even with basic models, simulation in R provides a clear and flexible way to explore potential outcomes and risks in financial markets.

**Palabras e frases chave**: Monte Carlo Simulation, Stochastic Calculus, Quantitative Finance, R (programming language).

## 1. INTRODUCIÓN

As **accións financeiras** son títulos de propiedade que representan unha parte proporcional do capital dunha empresa. Cando unha persoa ou institución merca unha acción, convértese en accionista e adquire dereito a participar nos beneficios (por exemplo, a través de dividendos) así como en determinadas decisións da compañía. O prezo dunha acción está determinado polo mercado, a través da oferta e da demanda, e recolle tanto o valor presente da empresa como as expectativas futuras sobre a súa evolución.

O estudo do comportamento das accións resulta fundamental en economía e finanzas, xa que constitúen un dos principais instrumentos de investimento. Non obstante, a súa dinámica é inherentemente incerta: factores económicos, políticos ou mesmo sociais poden afectar ao prezo dunha acción de maneira imprevisible. Esta natureza cambiante fai que resulte imposible predicir con exactitude o valor futuro dunha acción, e por iso xurdiu a necesidade de desenvolver modelos matemáticos que permiten aproximar e describir a súa evolución como un proceso aleatorio. Estes reciben o nome de **modelos estocásticos**.

Un dos modelos máis empregados é o chamado *movemento browniano xeométrico,* que vén dado pola solución da seguinte Ecuación Diferencial Estocástica (EDE):

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t$$

Obtendo como solución:

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right), \quad t \geq 0$$

Aquí $\sigma$ representa a volatilidade e $\mu$ o rendemento medio, un parámetro que reflicte a tendencia xeral do activo. Cómpre sinalar que tanto $\mu$ como $\sigma$ estímanse a partir dos datos históricos da acción, o cal nos dá unha idea razoable do comportamento típico do activo, mais tamén ten limitacións: nada garante que os patróns do pasado se manteñan no futuro, e de feito é moi probable que varíen co tempo.

Neste documento non se pretende falar sobre a importancia desta ecuación diferencial, mais é preciso mencionala xa que é a base teórica sobre a que se fundamentarán as simulacións que realizaremos a continuación.

## 2. METODOLOXÍA

O obxectivo central deste traballo é construír posibles escenarios de evolución para o prezo dalgún activo financeiro ao longo do tempo, tomando como referencia no noso caso os prezos das accións do banco BBVA. Para isto, seguiremos os seguintes pasos:

1. **Acceso ás bases de datos e extracción de parámetros**: empregamos a libraría `quantmod` en `R` para descargar desde *Yahoo Finance* as series históricas de prezos de peche axustados das accións de BBVA . Estes datos constitúen o punto de partida para a análise, xa que a súa evolución desde inicios de 2024 servirá de base para a estimación de $\mu$ e $\sigma$.

2. **Construción de camiños simulados**: a través da implementación numérica da EDE mediante pequenos incrementos, xeramos múltiples traxectorias aleatorias para o prezo da acción, onde cada camiño representa unha posible evolución distinta do activo ao longo do ano.

3. **Visualización e análise**: representamos graficamente algunhas das traxectorias. A partir delas, extráese información relevante coma o prezo esperado ou cuantís destacados, que serán tratados no seguinte apartado.

Todo este proceso permítenos analizar a información histórica dispoñible para obter unha imaxe clara da incerteza futura, facendo visibles tanto as oportunidades coma os riscos que implica investir nun activo como BBVA.

## 3. ANÁLISE DA SIMULACIÓN

Na Figura 1 obsérvanse múltiples traxectorias simuladas para o prezo das accións de BBVA ao longo dun ano bursátil real (252 días). Sobre a gráfica superpóñense as seguintes curvas:

- A esperanza teórica (azul discontinua): é o prezo esperado da acción en cada intre $t$, é dicir, a curva que vén dada pola seguinte función: $\mathbb{E}[S_t] = S_0 e^{\mu t}$

- A media empírica (vermella): é a media mostral do prezo da acción entre tódalas simulacións para cada instante $t$. Como podemos observar, atópase moi preto do valor teórico esperado, o que apunta a unha correcta execución do método.

- A mediana empírica (verde): liña que separa o 50 % dos resultados máis baixos dos máis altos. É interesante destacar que a curva da mediana sitúase lixeiramente por debaixo da media, xa que o valor elevado das simulacións máis optimistas non inflúe no cálculo da mediana.

- Value at Risk (morada discontinua): un concepto relevante en xestión de riscos que explicaremos a continuación.

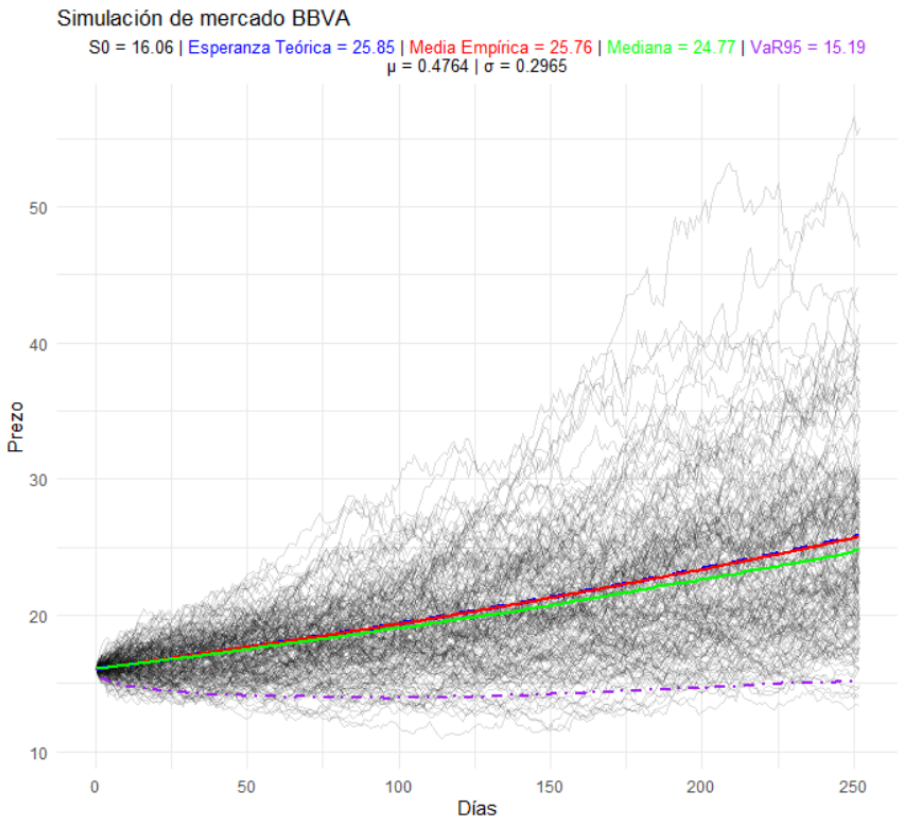



Figura 1: Evolución simulada dos prezos de BBVA

Desde o punto de vista financeiro, o **Value at Risk 5 %** (VaR), é o nivel a partir do cal a probabilidade de que activo baixe máis dese umbral nun tempo determinado é do 5 %. Ao realizar simulacións, podemos calcular o VaR de forma empírica como a porcentaxe de simulacións por debaixo dun certo rango, que cun gran número de simulacións se aproximará bastante ao valor real. Na gráfica a liña morada representa o VaR ao 5 %, é dicir, existe un 5 % de probabilidade de que as accións caian por debaixo desa liña.

Este resultado ofrece unha interpretación clara: non se trata de predicir un valor exacto para o futuro, senón de sinalar un rango dentro do cal é moi probable que se atope o prezo da acción. A visualización da curva do VaR xunto cos camiños simulados proporciona unha imaxe intuitiva da incerteza e do risco, convertendo unha idea estatística abstracta nunha ferramenta de análise accesible e útil.

Cómpre sinalar ademais que existen casos onde coñecer a distribución subxacente da solución da EDE faise virtualmente imposible, sexa polo complicado custo computacional ou mesmo pola imposibilidade de resolver a ecuación de maneira teórica. Nestes casos, a posibilidade de simular múltiples traxectorias como foi explicado neste documento, atópase entre as solucións máis eficaces, debido á robustez do método e á súa simplicidade.

## 4. CONCLUSIÓNS

O traballo presentado mostra como é posible empregar modelos estocásticos sinxelos para aproximar o comportamento das accións financeiras e xerar escenarios de futuro. A través de `R` conseguimos integrar nun mesmo contorno a descarga de datos reais e a simulación de prezos de mercado.

Como xa se estableceu anteriormente, a virtude deste enfoque non reside en ofrecer predicións exactas, senón en proporcionar unha ferramenta flexible que permite explorar a incerteza e comprender mellor o risco inherente aos mercados financeiros. As representacións gráficas derivadas das simulacións facilitan transmitir conceptos complexos dun xeito visual e accesible, contribuíndo a que o estudo dos mercados sexa máis intuitivo.

# Referencias


[1] I. Kharroubi, *Calcul stochastique et introduction au contrôle stochastique, et applications à la finance*. Polycopié de Master 1, Sorbonne Université, 2025. (D'après un polycopié original de Philippe Bougerol).

[2] R. J. Shiller, *Financial Markets*. Curso en liña ofrecido por Yale University a través de Coursera, 2017. Dispoñible en: `https://www.coursera.org/learn/financial-markets-global`.

[3] Yahoo Finance, *Datos históricos de mercado para BBVA e outros activos*. Disponible en: `https://finance.yahoo.com`.

[4] R. Jahn e colaboradores, *quantmod: Quantitative Financial Modelling Framework*. R package, CRAN. Dispoñible en: `https://cran.r-project.org/package=quantmod`.

[5] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, 2016. Paquete de R dispoñible en: `https://cran.r-project.org/package=ggplot2`.

[6] H. Wickham, *reshape2: Flexibly Reshape Data*. R package, CRAN. Dispoñible en: `https://cran.r-project.org/package=reshape2`.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Análise do efecto da consulta electrónica no Servizo de Cardioloxía da Área Sanitaria de Santiago de Compostela e Barbanza

Iria Iglesias Gende[1], Mercedes Conde Amboage[2,3] e Francisco Reyes Santías[4,5]

[1]Biostatech, Advice, Training & Innovation in Biostatistics, S.L.
[2]Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.
[3]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga).
[4]Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS).
[5]Departamento de Organización de Empresas e Mercadotecnia. Universidade de Vigo.

## RESUMO

Neste traballo estudamos o impacto da introdución da consulta electrónica (habitualmente coñecida como e-consulta) no Servizo de Cardioloxía da Área Sanitaria de Santiago de Compostela e Barbanza sobre o tempo de vida das/os pacientes derivados a dito servizo.

Dado que a base de datos proporcionada pola Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS) presenta información censurada, empregamos técnicas de Análise de Supervivencia para analizar a influencia da introdución da e-consulta no tempo de vida das/os pacientes. En particular, utilizamos modelos aditivos xeneralizados (coñecidos habitualmente polas súas siglas en inglés: GAM) incorporando os pesos de Kaplan-Meier, co obxectivo de adaptar a metodoloxía clásica ao contexto dos datos censurados e estimar de maneira flexible a relación entre as diferentes variables explicativas consideradas e o tempo de vida. Este enfoque permite avaliar ata que punto a implantación da e-consulta pode repercutir na mellora dos tempos de supervivencia das/os pacientes e, en consecuencia, na calidade da atención sanitaria.

Para levar a cabo esta análise, imos empregar o software estatístico R xunto con diferentes librarías tales como `survival`, `condSURV`, `mgcv` ou `gratia`.

**Palabras e frases chave**: e-consulta, cardioloxía, análise de supervivencia, GAM, Kaplan-Meier.

## 1. INTRODUCIÓN

O desenvolvemento da historia clínica electrónica integrada entre niveis asistenciais, coñecida en Galicia como IANUS, supuxo unha transformación significativa na xestión clínica ambulatoria. Este sistema permite o acceso compartido á información sanitaria por parte das/os profesionais de atención primaria e especializada, favorecendo así a continuidade asistencial e a coordinación entre niveis. Ademais, garante a accesibilidade á totalidade da historia clínica da/o paciente, con independencia do centro sanitario no que sexa atendida/o, e contribúe á homoxeneidade na incorporación e rexistro da información clínica.

O Servizo de Cardioloxía da Área Sanitaria de Santiago de Compostela e Barbanza implantou as consultas de acto único no ano 2008. Este modelo asistencial permitía realizar na mesma xornada a consulta presencial e as probas diagnósticas necesarias, o que permitía reducir a asistencia a un só acto médico, fronte aos modelos máis clásicos que incluían diferentes visitas da/o paciente para a consulta. Posteriormente no ano 2013, púxose en marcha un proxecto de e-consulta consensuado entre as/os profesionais, que inclúe unha primeira consulta non presencial, isto é a e-consulta,

grazas á existencia da IANUS. A través desta modalidade, as/os diferentes profesionais sanitarios poden remitir unha interconsulta especificando o motivo da demanda, e a/o cardióloga/o responde baseándose nos datos clínicos dispoñibles no historial da/o paciente. Se a información é suficiente establécese unha estratexia diagnóstica sen necesidade de cita presencial; en caso contrario, a/o paciente é derivada/o a unha consulta presencial de acto único na que se toma a decisión da alta clínica ou do seguimento en consultas posteriores por parte do Servizo de Cardioloxía.
Á vista da situación exposta, realizamos unha análise do efecto da implantación da e-consulta no tempo de vida grazas a unha base de datos proporcionada pola Fundación Instituto de Investigación Sanitaria (FIDIS) que contén o seguimento de 75763 pacientes atendidas/os polo Servizo de Cardioloxía da Área Sanitaria de Santiago de Compostela e Barbanza entre os anos 2008 e 2023.

## 2. METODOLOXÍA E RESULTADOS

Debido á presenza de datos censurados na base de datos proporcionada, empregamos técnicas de Análise de Supervivencia, que é a rama da Estatística que estuda o tempo que transcorre ata que un certo conxunto de individuos experimente un certo evento de interese.
Unha característica distintiva nesta base de datos é a censura pola dereita: o evento final non se pode determinar con certeza para todos os individuos, posto que o tempo de vida só pode ser observado parcialmente. Debido a isto, o estimador empírico tradicional deixa de ser consistente, polo que se xeneralizou ao caso de datos censurados, dando lugar ao coñecido como estimador de Kaplan-Meier, ver [3], que se define como:

$$\widehat{F}(t) = 1 - \prod_{T_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{1}$$

onde $T_1 < ... < T_j$ son os tempos observados da ocurrencia do fallo (é dicir, son os $Z_i$ tales que $\delta_i = 1$), $d_i$ o número de fallos en $T_i$ e $n_i$ o número de individuos que están a risco en $T_i$, é dicir, que non morreron nin foron censurados antes de $T_i$.
Axustando o estimador de Kaplan-Meier definido en 1, podemos obter as curvas de supervivencia estimadas para os dous tipos de consulta considerados (a consulta de acto único e a e-consulta). Estas curvas mostran que a función de supervivencia tende a ser maior no grupo de pacientes derivadas/os a e-consulta. De todos xeitos, para avaliar formalmente esta diferenza, utilizamos o test *log-rank*, cuxas hipóteses son as seguintes:

$$\begin{cases} H_0 : h_1(t) = h_2(t), \ t \geq 0, \\ H_1 : h_1(t) \neq h_2(t), \ \text{para algún } t, \end{cases} \tag{2}$$

onde $h_i$ representa a función de risco, isto é, a probabilidade de que, se un individuo sobrevive no tempo $t$, experimente o fallo no instante inmediato despois $[t, t + dt)$.
Grazas ao contraste 2, observamos que existen diferenzas estatisticamente significativas entre os dous grupos, reforzando a evidencia visual obtida co estimador de Kaplan-Meier de que a introdución da e-consulta aumenta o tempo de vida das/os pacientes.
Estes resultados motivaron a búsqueda de modelos de regresión que analicen máis en detalle como evolucionan os tempos de supervivencia segundo as diferentes covariables que hai dispoñibles na base de datos. En particular, centrámonos na proposta de Orbe e Virto, ver [4], que adapta os modelos aditivos xeneralizados, ver [2] e [6], para o contexto de datos censurados.
Dita proposta combina o emprego de bases de funcións P-*splines* cun método de estimación baseado en mínimos cadrados ponderados, co obxectivo de ter en conta que estamos a traballar con datos en presenza de censura. Trátase dunha extensión da proposta de Eilers e Marx, ver [1], que se baseaba no emprego de P-*splines* pero que integra os pesos Kaplan-Meier, seguindo a idea da proposta de Stute, ver [5], para manexar adecuadamente a presenza de observacións censuradas.
Consideramos entón o seguinte modelo:

$$Y_i = \alpha + \sum_{j=1}^{p} m_j(X_{ij}), \tag{3}$$

onde $Y_i = \log(T_i)$ é o logaritmo do tempo de supervivencia do individuo $i$ e $m_j$ representa o efecto non paramétrico de cada variable explicativa sobre a variable resposta.

Os resultados do modelo 3 evidenciaron unha asociación positiva entre a presenza da e-consulta e un maior tempo de vida das/os pacientes. Deste xeito, os resultados evidencian que a atención mediante a e-consulta contribúe positivamente aos resultados en saúde, posiblemente debido a unha atención máis precoz que evita que as/os pacientes acaden situacións críticas.

### 3. CONCLUSIÓNS

A través do enfoque presentado, puidemos observar unha asociación positiva entre a presenza da e-consulta e un maior tempo de vida das/os pacientes. Este resultado é especialmente relevante se se ten en conta que a e-consulta non implica necesariamente unha intervención clínica directa, senón que actúa como un mecanismo de coordinación e triaxe máis eficiente entre os diferentes niveis asistenciais. Ademais, estes achados reforzan a idea de que a transformación dixital do sistema sanitario non só afecta á eficiencia organizativa, senón que tamén ten efectos positivos na calidade asistencial e no benestar da poboación.

## Referencias


[1] Eilers, P.H.C. e Marx, B.D. (1996). *Flexible smoothing with B-splines and penalties.* Statistical Science, 11(2), 89-121.

[2] Hastie T.J. e Tibshirani, R.J. (1990). *Generalized additive models.* Chapman and Hall.

[3] Kaplan, E.L. e Meier, P. (1958). *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, 53(282), 457–481

[4] Orbe J. e Virto, J. (2018). *Penalized spline smoothing using Kaplan–Meier weights with censored data.* Biometrical Journal, 60(5), 947-961.

[5] Stute, W. (1993). *Consistent estimation under random censorship when covariables are present.* Journal of Multivariate Analysis, 45(1), 89–103.

[6] Wood, S.N. (2017). *Generalized additive models. An introduction with R* 2nd Edition. CRC Press.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# R COMO HERRAMIENTA PARA EL ESTUDIO DE PATRONES BIOGEOGRÁFICOS

Ramiro Martín-Devasa[1], Sara Martínez-Santalla[2], Carola Gómez-Rodríguez[3], Rosa M. Crujeiras[4] & Andrés Baselga[2]

[1]Department of Geosciences and Geography, University of Helsinki.

[2]CRETUS, Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela.

[3]CRETUS, Department of Functional Biology (Area of Ecology), Universidade de Santiago de Compostela, Santiago de Compostela.

[4]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Santiago de Compostela.

## RESUMEN

Una gran variedad de patrones ecológicos se modelizan usando funciones paramétricas. Estas son muy útiles, ya que permiten cuantificar y comparar los diferentes patrones a través del valor de sus parámetros, y estimar los efectos de las variables ambientales y geográficas sobre diferentes especies, poblaciones o comunidades biológicas. Sin embargo, no todas las funciones son válidas para todos los patrones, y la estructura de dependencia a pares de muchas de las variables (e.g. distancias geográficas o ambientales) dificultan las comparaciones entre ajustes. Aquí proponemos nuevos modelos para el ajuste de patrones biogeográficos y nuevos métodos para evaluar su nivel de significación y realizar comparaciones de parámetros en de funciones ajustadas con variables con dependencia a pares. Todos estos métodos han sido implementados en el paquete `betapart` de R.

**Palabras y frases clave**: *Distance-decay*, función paramétrica, patrones biogeográficos, dependencia a pares.

## 1. INTRODUCCIÓN

Una de las cuestiones más estudiadas en ecología, y especialmente en biogeografía son las causas del cambio en composición de especies entre diferentes lugares. ¿Qué causa que diferentes localidades, más o menos cercanas geográficamente, tengan diferentes especies de animales, plantas, etc.? Una de los métodos más utilizados para estudiar esta cuestión se basa en estimar la similitud de composición de especies entre diferentes lugares, i.e. cómo de diferentes son sus comunidades

biológicas, y modelizar mediante funciones paramétricas la disminución de esta similitud con la distancia geográfica y ambiental entre las localidades estudiadas. Este patrón de disminución de la similitud con la distancia espacial se conoce como *distance-decay*. Históricamente, el *distance-decay* se ha estudiado mediante funciones *power-law* y exponencial negativa. Sin embargo, estas funciones son incapaces de modelar correctamente situaciones en las que las comunidades cercanas son prácticamente idénticas y la similitud no decae hasta media o larga distancia. Para recoger correctamente este tipo de relaciones, una función sigmoide, que tuviera en cuenta similitudes constantes a corta distancia, sería necesaria. Por ello en Martín-Devasa et al.(2022) se propuso la función de Gompertz como una nueva alternativa para el ajuste y estudio del *distance-decay*. Actualmente, la función *decay.model* del paquete de R `betapart` permite ajustar tanto modelos *power-law* y exponencial negativo como Gompertz, a datos de *distance-decay*, siendo la principal función en R para este tipo de análisis.

Tanto los índices de similitud entre comunidades como las medidas de distancia espacial y climática se calculan por la comparación entre dos puntos, por lo que presentan dependencia a pares. Esto dificulta calcular su nivel de significación y hacer comparaciones de parámetros entre modelos, ya que se produce una inflación de los grados de libertad que puede aumentar el error de tipo I en unos modelos que, a demás, son no lineales. Para poder evaluar el nivel de significación de modelos de *distance-decay*, Martínez-Santalla et al. (2022) propusieron un test de significación basado en remuestreo por bloques. Definiendo un bloque como el conjunto de todos los valores de similitud o distancia en las que participa el mismo punto, y tomando esos bloques como unidad de muestreo, se puede calcular la distribución nula de un test, en este caso de significación, sin que el efecto de la inflación de los grados de libertad. Como estadístico, se propuso el uso del pseudo-$R^2$, una medida similar al $R^2$, pero que es útil para modelos no lineales. Así, la hipótesis nula del test sería la independencia entre la similitud entre comunidades y su distancia espacial o ambiental (la composición de especies no cambia ni con la distancia espacial o con diferencias ambientales), y la hipótesis alternativa sería que hay un efecto de la distancia espacial entre localidades o de sus diferencias ambientales en la composición de sus especies. Este test también está implementado en la función *decay.model* de `betapart`.

Asimismo, el remuestreo por bloques utilizado para el test de significación se puede usar para construir un test para la comparación de parámetros en modelos construidos con datos con dependencia a pares. Martín-Devasa et al.(2022) propusieron un test para la comparación de parámetros de modelos de *distance-decay* llamado Z_dep. Este test (basado en la diferencia de parámetros) tiene la misma estructura que un test t para la diferencia de medias. La diferencia está que la variancia de la diferencia de parámetros se estima con un remuestro por bloques, evitando la inflación de los grados de libertad y limitando el error de tipo I. Este test está ahora disponible en la función *zdep* del paquete `betapart`.

## 2. R PARA EL MODELADO DEL *DISTANCE-DECAY* Y PATRONES BIOGEOGRÁFICOS

El paquete `betapart`, a demás de varias opciones para el cálculo de índices de similitud entre comunidades, ofrece una serie de funciones para modelar patrones de *distance-decay*. Como se ha adelantado en la introducción, la función *decay.model* es de gran utilidad para ajustar las tres funciones utilizadas en el modelado del *distance-decay*: *power-law*, exponencial negativa y Gompertz.

La función *zdep* es la principal alternativa para la comparación de parámetros entre modelos de *distance-decay*, y el test puede ser utilizado de forma general para la comparación de parámetros en modelos ajustados con datos con dependencia a pares.

# Referencias


[1] Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* 30(3), 279–338.

[2] Nekola, J. C., & White, P. S. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography* 26(4), 867–878.

[3] Nekola, J. C., & McGill, B. J. (2014). Scale dependency in the functional form of the distance decay relationship. *Ecography* 37(4), 309–320.

[4] Martín-Devasa, R., Martínez-Santalla, S., Gómez-Rodríguez, C., Crujeiras, R. M., Baselga, A. (2022). Comparing distance-decay parameters: A novel test under pairwise dependence. *Ecological Informatics* 72,101894.

[5] Martínez-Santalla, S., Martín-Devasa, R., Gómez-Rodríguez, C., Crujeiras, R.M., Baselga, A. (2022). Assessing the nonlinear decay of community similarity: permutation and site-block resampling significance tests. *Journal of Biogeography* 49(5), 968–978.

[6] Baselga, A., & Orme, C. D. L. (2012). Betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution* 3(5), 808–812.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Capturando Bordas de Grafismos Marajoaras para Construção de Mandalas Matemáticas

João Paulo Martins dos Santos[1] e Luciane Ferreira Alcoforado[2]

[1,2]Academia da Força Aérea Brasileira

## RESUMO

Os algoritmos de processamento de imagens foram utilizados em grafismos marajoaras para detecção de contornos e construção das mandalas matemáticas. O tema é um projeto de longo prazo de desenvolvimento e integração do R com elementos de Geometria e, mais geral, da Matemática. O processo de construção das mandalas matemáticas segue os aspectos delineados em (Alcoforado *et al.*; 2023). As artes marajoaras foram escolhidas devido à riqueza cultural combinado ao interesse dos autores em explorar diferentes formas geométricas para a construção das mandalas matemáticas. A metodologia de processamento de imagens para a detecção de contornos foi baseada nos pacotes `imager` e `magick`, os quais combinados com os pacotes `dplyr` e `ggplot2` fornecem os elementos para o processamento, análise e visualização da arte de interesse. A modelagem proposta é um avanço em relação aos métodos anteriores em que curvas planas com interpretações ou releituras foram utilizados tornando o processo mais preciso. Os resultados apontam para uma diversificação dos métodos de construção de mandalas matemáticas, traduzidos em um extenso número de combinações de formas e cores possibilitados por uma única peça de arte. Ao utilizar a linguagem R, os elementos de Matemática, Computação e linguagem de programação de alto nível são combinados aos elementos artísticos e culturais para a expansão dos horizontes de possibilidades. Espera-se que os resultados despertem interesse em novas pesquisas relacionadas ao tema e insiram o R ainda mais no contexto das aplicações matemáticas.

**Palabras e frases chave**: Processamento de Imagens, Detecção de Bordas, Grafismos.

## 1. INTRODUÇÃO

Os grafismos marajoaras são sistemas visuais amazônicos caracterizados por padronagens geométricas que combinam simetria, abstração e simbolismos. Segundo Schann (2008), uma das características mais marcantes da cerâmica marajoara é o convívio, em um mesmo objeto, de representações naturalistas e representações geometrizantes, estas últimas chamadas usualmente de grafismos. Estes grafismos podem ser utilizados em um contexto pedagógico para

O presente artigo propõe uma metodologia para a construção de mandalas matemáticas baseadas na extração de contornos das imagens de grafismos da cultura marajoara. A digitalização e análise computacional desses elementos visuais contribuem para uma releitura dos elementos culturais, por meio de aproximações das curvas, traçados e representações dos objetos. Essa aproximação de traçados por meio de uma técnica computacional permite a releitura, reconfiguração e ressignificação dos elementos culturais em um novo contexto original com a finalidade de expansão dos horizontes dessa arte.

## 2. METODOLOGIA

A linguagem R será utilizada para a identificação e aproximação de bordas em imagens de grafismos marajoaras por meio do pacote `imager` (Barthelme, 2024), `magick` (Ooms, 2024), `dplyr` (Wickham *et al.*, 2023) e `ggplot2` (Whickham, 2016).

A metodologia para processamento de imagens combina abordagens complementares dos pacotes `magick` e `imager`. Para o pacote `imager` a metodologia consiste em:i.) Pré-processamento: carregar imagem e tratar transparência. Converter para RGB e escala de cinza; ii.) Binarização: aplicar *threshold* para criar imagem preto e branco; iii.) Extração: identificar fronteiras entre regiões e armazenar coordenadas. Por sua vez, a metodologia do pacote `magick` compreende: i) conversão da imagem para escala de cinza; ii) binarização mediante *threshold* adaptativo; iii) detecção de bordas utilizando algoritmo de Canny; iv) extração de coordenadas de contorno.

Em ambos os casos, a etapa de pós-processamento consiste em: iv.) Pós-processamento: filtrar contornos de acordo com um parâmetro relacionado ao número de pontos; v.) Análise: inspeção visual para remover contornos indesejados e visualização final; vi.) Seleção: seleção de contornos desejados e construção das Mandalas: utilização dos contornos (denominados níveis), únicos ou composições, como módulo básico e aplicação de transformações geométricas (translações, rotações, reflexões e homotetias) para gerar o padrão mandálico simétrico; vii.) Pós-processamento e Análise: refinamento por meio de uma paleta de cores para aplicação às mandalas geradas. Os códigos utilizados para a geração das mandalas das Figuras 3 e 4 podem ser obtidos em GitHub Link.

A Figura 1 ilustra os contornos identificados utilizando os pacotes `imager` e `magick` para a imagem de um prato marajoara.


(a) Imagem final de a.) após filtro de eliminação de contornos com menos de 400 pontos.

(b) Imagem final de b.) após filtro de eliminação de contornos com menos de 400 pontos.

Figura 1: Bordas detectadas com o R utilizando os pacotes `imager` e `magick`: diferenciação dos níveis por meio das cores.

Em ambos os casos, os diferentes contornos, representados pela variável *nível*, são identificados pelas diferentes cores. No caso do pacote `imager`, os níveis são fechados e adequadamente identificados, enquanto os níveis identificados pelo pacote `magick` não fornecem estruturas fechadas. Entretanto, os níveis individuais ou combinados de ambos podem ser utilizados para a construção das mandalas matemáticas conforme resultados nas Figuras 3 e 4.

As paletas de cores utilizadas nas composições das Figuras 3 e 4 são mostradas, respectivamente, na Figura 2. Essas cores foram obtidas por meio de um processo imaginativo de releitura dos autores. Tal processo pode ser replicado em outros contextos ou com outros grafismos.

Figura 2: Paleta de cores para as Figuras 3 e 4, respectivamente.

O código a seguir, utilizando o algoritmo de *Canny*, é um exemplo que permite a extração dos contornos com filtragem dos níveis com menos de 400 pontos. O valor $N = 400$ foi determinado por análise direta dos resultados na Figura 1, enquanto que o valor e $centers = 20$ foi determinado com base nos resultados do pacote `imager`.

```
library(magick); library(ggplot2); library(dplyr)
file <- getwd(); input_image_path <- file.path(file, "marajo2.jpg")
img_magick <- image_read(input_image_path)
img_gray <- image_convert(img_magick, colorspace = "Gray")
```

```
edges <- image_canny(img_gray, geometry = "0x1+10%+30%")
image_write(edges, "bordas_canny.png"); edges_imager <- load.image("bordas_canny.png")
edges_imager <- load.image("bordas_canny.png")
par(mfrow = c(1, 2)); plot(img_gray, main = "Imagem Original (Cinza)")
plot(edges_imager, main = "Bordas Detectadas (Canny)")
pontos_contorno <- which(edges_imager > 0.5, arr.ind = TRUE)
final_df <- data.frame( x = pontos_contorno[, 1],y = pontos_contorno[, 2])
set.seed(123); clusters <- kmeans(final_df[, c("x", "y")], centers = 20)
final_df$level <- clusters$cluster
N <- 400; final_df_filtrado <- final_df %>%group_by(level) %>%filter(n() >= N) %>%
  ungroup() %>% mutate(level = as.numeric(factor(level)))
```

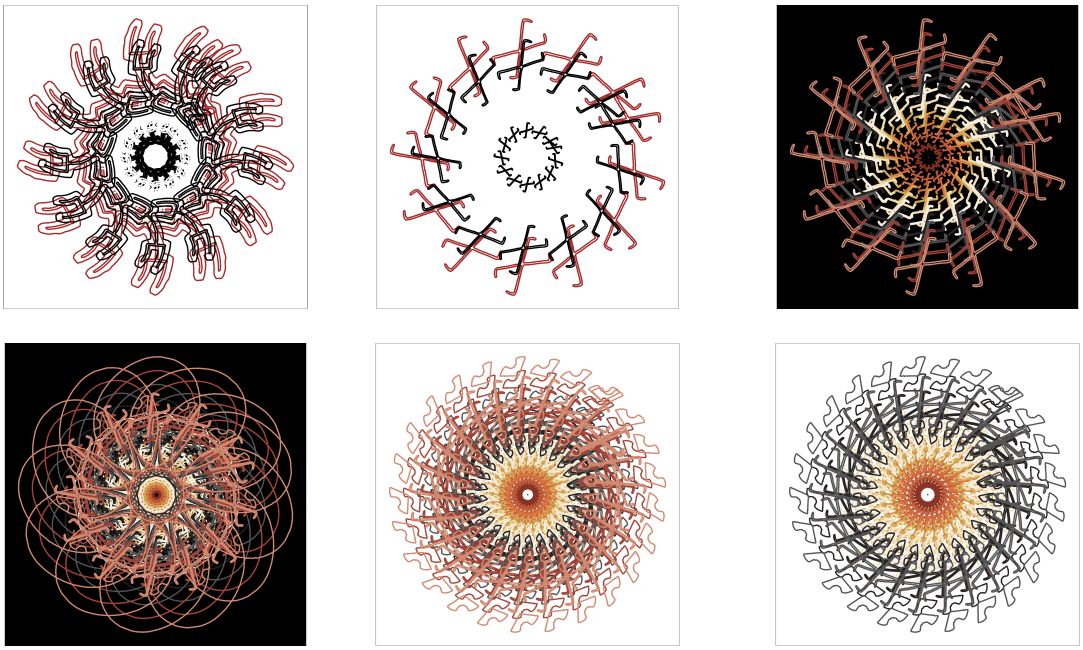


Figura 3: Mandalas Matemáticas utilizando contornos (níveis) ou composições de níveis da Figura 1a. Fonte: Autores, 2025.

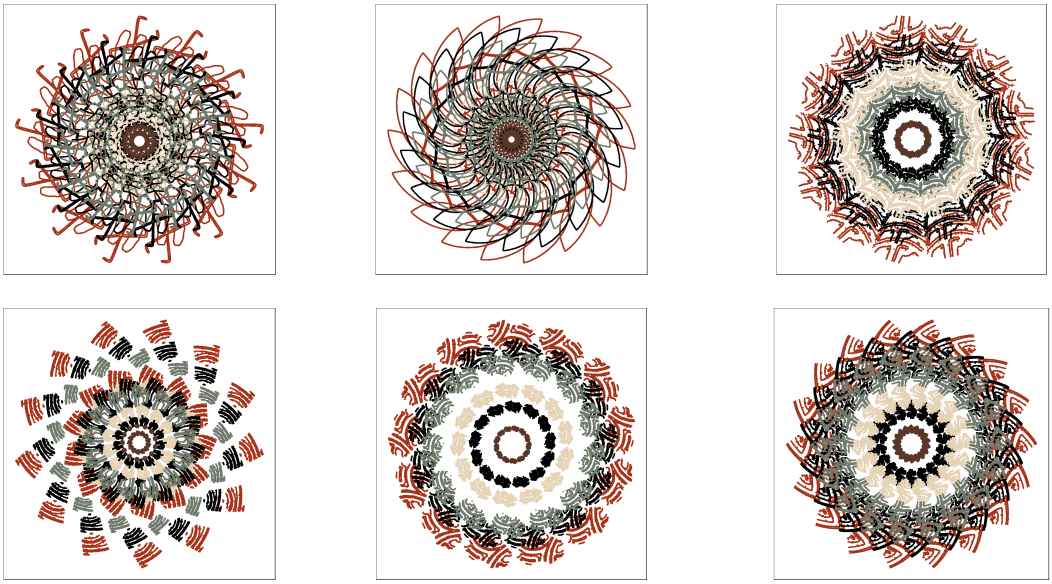


Figura 4: Mandalas Matemáticas utilizando contornos (níveis) ou composições de níveis da Figura 1b. Fonte: Autores, 2025.

As Figuras 3 e 4 refletem o processo de construção utilizando apenas algumas curvas dentre o conjunto de possibilidades unitárias ou combinações. A combinação de níveis disponíveis nos resultados obtidos por meio dos pacotes pode adicionar possibilidades ao extenso conjunto de construções. Deve ser notado que o processo de construção mantém o viés adotado em Alcoforado *et al.* (2023), o qual restringe a construção das mandalas à utilização de rotações, homotetias e translações, com as rotações no intervalo $[0, 2\pi]$ sendo predominantes.

## 4. CONCLUSÕES

A construção de mandalas matemáticas não é um assunto inédito, no entanto, o contexto da extração de contornos de figuras originadas de elementos culturais e artísticos, como por exemplo, dos grafismos marajoaras insere novas possibilidades e expande os horizontes de integração entre diferentes campos do conhecimento. Os pacotes `imager` e `magick` são duas possibilidades de utilização da linguagem R para fins de detecção de contornos e, consequente utilização na construção de mandalas matemáticas. Espera-se que os resultados possam servir da base para novas investigações.


### AGRADECIMENTOS

À Academia da Força Aérea e aos professores, distintos colegas, do Grupo de Modelagem Matemática e Computacional (GMMC).


# Referencias


[1] Alcoforado, Luciane Ferreira; Santos, João Paulo Martins dos; Lima, Marcus Vinícius de Araújo; Jesus, Alessandro Firmiano de & Linares, Juan López. (2023). **Mandalas, curvas clássicas e visualização com R**. Universidade de São Paulo. Faculdade de Zootecnia e Engenharia de Alimentos. DOI: https://doi.org/10.11606/9786587023335 Disponível em: www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/1017 . Acesso em 21 Set. 2025.

[2] Barthelme, Simon. (2024) *imager: Image Processing Library Based on 'CImg'*. R package version 1.0.2. Disponível em: `https://CRAN.R-project.org/package=imager`.

[3] Ooms, Jeroen. (2024) *magick: Advanced Graphics and Image-Processing in R*. R package version 2.8.4. Disponível em: `https://CRAN.R-project.org/package=magick`.

[4] Schaan, D. P. (2008). A Arte da Cerâmica Marajoara: encontros entre o passado e o presente. **Revista Habitus - Revista do Instituto Goiano de Pré-História e Antropologia**, Goiânia, Brasil, v. 5, n. 1, p. 99–117. DOI: 10.18224/hab.v5.1.2007.99-117. Disponível em: https://seer.pucgoias.edu.br/index.php/habitus/article/view/380. Acesso em: 21 set. 2025.

[5] Wickham, Hadley; François, Romain; Henry, Lionel; Muller, Kirill; Vaughan, Davis. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4. Disponível em: `https://CRAN.R-project.org/package=dplyr`.

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

*XII Xornada de Usuarios de R en Galicia*

*Santiago de Compostela, 16 de outubro do 2025*

# Genómica de la conservación de la especie *Daphne gnidium* en poblaciones costeras atlánticas de Galicia.

Carlos Mirás-Viña[1], Adrián Casanova[1], Fernando Cabana[1], Miguel Hermida[1], Andrés Blanco[1], Javier Ferreiro[2], Pablo Ramil[2], Carmen Bouza[1], Manuel Vera[1]
carlosmirasvina@gmail.com

1. Grupo de investigación ACUIGEN (GI-1251), Departamento de Zoología, Genética y Antropología Física, Facultad de Veterinaria, Campus Terra, Universidade de Santiago de Compostela, 27002, Lugo, España, mcarmen.bouza@usc.es, manuel.vera@usc.es
2. GI-1934-TB, IBADER, Campus Terra, Universidade de Santiago de Compostela, 27002, Lugo, España, ramil.rego@usc.es

## RESUMEN

Este trabajo aborda la conservación genómica de la especie *Daphne gnidium* en poblaciones costeras atlánticas de Galicia mediante análisis de diversidad genética, estructura poblacional y variación adaptativa, utilizando datos obtenidos por secuenciación 2b-RADseq. Se emplearon diferentes herramientas bioinformáticas libres y gratuitas, como R a través de RStudio para el procesamiento, filtrado, análisis estadístico y visualización de datos genómicos. La integración de diversos paquetes en R permitió la identificación de SNPs robustos y estructuras genéticas relevantes para la conservación genética y recomendaciones para la gestión in situ de la especie.

**Palabras clave:** *Daphne gnidium*, secuenciación 2b-RADseq, SNPs, genómica de conservación, diversidad genética, estructura poblacional.

## 1. INTRODUCCIÓN

*Daphne gnidium* (Fig. 1) es un arbusto de tipo esclerófilo denominado torvisco. Se trata de una especie diploide (2n=18) con capacidad de reproducción sexual y elevada capacidad de revegetación, distribuyéndose en zonas costeras protegidas.
La especie presenta un gran interés tanto medicinal como de conservación. Su importancia medicinal radica en la presencia de compuestos fitoquímicos en aceites esenciales y extractos, principalmente de las hojas, responsables de sus propiedades biológicas y terapéuticas (antioxidantes, antibacterianos, neuroprotectores y antitumorales). Además, *D. gnidium* es una especie clave para la conservación de las dunas costeras, sus raíces profundas ayudan a estabilizar y fijar estas estructuras, evitando su erosión. Asimismo, la planta proporciona hábitat y recursos a una gran variedad de especies que habitan en estos ecosistemas.
Para este estudio se han utilizado SNPs como marcadores moleculares. la irrupción de técnicas de secuenciación masiva de lecturas cortas (*short reads*, e.g., Illumina) y largas (*long reads,* e.g., Oxford Nanopore™) y la reducción de su coste por megabase ha facilitado su aplicación en aproximaciones genómicas. Estas técnicas genómicas han facilitado la detección y genotipado de grandes paneles de SNPs. Trabajar con miles o

millones de marcadores proporciona nuevos conocimientos y comprensión de procesos en el campo de la conservación.

La evaluación de la diversidad genética es fundamental para garantizar la conservación efectiva de las especies. La información genómica constituye una herramienta clave para diseñar acciones concretas de conservación y gestión, adaptadas a las necesidades de cada especie. Por ello, el proyecto LIFE INSULAR (https://www.lifeinsular.eu/) se enfoca en la restauración de hábitats y la protección de especies clave de los mismos como *D. gnidium*, actualmente amenazada por la degradación de las dunas.

A B

**Figura 1. A:** Bayas de *Daphne gnidium L.* Fuente: © Pablo Vargas | RJB-CSIC. **B:** Planta florada de *D gnidium* L. Fuente: IBADER-USC (https://www.ibader.gal/).

## 2. OBJETIVOS

Los objetivos principales fueron evaluar la calidad de los datos brutos de genotipado generado por la metodología 2b-RADseq, desarrollar un panel robusto de SNPs, y analizar la diversidad genética y estructura poblacional de *Daphne gnidium* en seis localidades costeras atlánticas protegidas de Galicia. Todo ello para la obtención de información que pueda ser aplicable al apoyo de la gestión y conservación de recursos genéticos de la especie en las áreas a estudio.

## 3. METODOLOGÍA

### 3.1 Muestreo y extracción ADN

Se recogieron muestras de seis localidades, dentro de áreas protegidas como Zonas Especiales de Conservación (ZEC) y Parque Nacional Illas Atlánticas (PNIA; Tabla 1). La extracción de ADN se realizó siguiendo protocolos estándar para asegurar una calidad apta para secuenciación masiva.

**Tabla 1.** Localidades de muestreo, código genético, coordenadas (WGS84), número de ejemplares (N) y áreas protegidas (ZEC/PNIA) de la especie Daphne gnidium analizados en este estudio. * ZEC Red Natura (Zonas Especial Conservación) / Parque Nacional Illas Atlánticas (PNIA).

| Localidad | Código genético (Coordenadas, X; Y) | N | Áreas protegidas* |
|---|---|---|---|
| Barra | C1 (-8,836; 42,263) | 19 | ZEC-Costa da Vela |
| O Grove | C2 (-8,895; 42,473) | 16 | ZEC-Complejo Ons-Grove |
| Corrubedo | C3 (-9,028; 42,549) | 16 | ZEC-Complejo Corrubedo |
| Cíes | I1 (-8,899; 42,228) | 19 | ZEC Illas Cíes / PNIA |
| Ons | I2 (-8,928; 42,387) | 18 | ZEC Complejo Ons-Grove / PNIA |
| Sálvora | I3 (-9,004; 42,470) | 20 | ZEC Complejo Corrubedo / PNIA |
| Total | | 108 | |

### 3.2 Secuenciación y procesamiento bioinformático

El genotipado se llevó a cabo mediante la metodología 2b-RADseq con Illumina NextSeq 500, utilizando enzima de restricción *Alf I*. La calidad de datos se evaluó con

FastQC 0.12.0, y el resumen global se integró con MultiQC. Para el alineamiento, se aplicó la herramienta Bowtie1.3.1 con los genoma de referencia de especies congenéricas y un enfoque *de novo* con Stacks 2.64 para identificar y filtrar SNPs.

### 3.3 Caracterización genética poblacional

Se realizó el análisis de los genotipos, de diversidad genética y estructura de los paneles de SNPs obtenidos con genoma de referencia y *de novo*. A partir del panel completo *de novo* se generó un panel de SNPs *outliers* divergentes (loci o regiones del genoma que puedan estar bajo selección) y se realizó el análisis de diferenciación y estructura genética.

RStudio fue la plataforma central para el análisis estadístico y visualización de datos. Se emplearon paquetes especializados para analizar diversidad genética, estructura poblacional y diferenciación:

- **Genepop 1.2.2**: Para edición, filtrado y análisis de datos genéticos multilocus en formato GENEPOP. Se usó para calcular frecuencias alélicas, heterocigosis observada y esperada y realizar análisis de equilibrio Hardy-Weinberg y diferenciación genética ($F_{ST}$).
- **DiveRsity 1.9.90**: Para estimar parámetros de diversidad genética poblacional (la materia prima de la evolución), incluyendo número de alelos, riqueza alélica corregida por rarefacción, heterocigosis, índice de fijación ($F_{IS}$) con intervalos de confianza bayesianos.
- **StAMPP 1.6.3**: Para el cálculo de diferenciación genética y estructura poblacional, especialmente para estimar el coeficiente $F_{ST}$ y sus intervalos de confianza mediante *bootstrapping* entre loci.
- **Adegenet 2.1.11**: Para manipulación y análisis de datos genéticos multilocus en objetos de R de clase genind/genlight, realizando análisis multivariantes como el Análisis Discriminante de Componentes Principales (DAPC) para identificar agrupaciones genéticas y visualizar su estructura.
- **dartR 2.9.9.5**: Para la conversión y manejo de datos genotípicos SNP en formatos compatibles, facilitando la interoperabilidad entre diferentes análisis y formatos de datos.
- **ParallelStructure 1.0.0**: Para ejecutar en paralelo el programa STRUCTURE en múltiples núcleos de procesador desde R, optimizando el análisis bayesiano de agrupamiento genético para determinar el número de unidades poblacionales genéticamente homogéneas (K).

La combinación de estos paquetes permitió identificar patrones de estructura local entre pares isla-continente y trabajar con SNPs con posibles señales de selección adaptativa.

## 4. RESULTADOS

Se obtuvieron aproximadamente 537 millones de lecturas de ADN, con una tasa de retención del 81,4% tras el filtrado. El panel generado por alineamiento *de novo* se destacó por su robustez y representatividad. Los análisis de diversidad y diferenciación genética se centraron en el panel *de novo*.

Los análisis mostraron una diversidad genética moderada, similar entre localidades. Hubo un bajo tamaño efectivo ($N_e < 50$; Fig. 2). Se detectó una estructuración genética clara entre las poblaciones insulares y continentales (p.ej., Cíes-Barra, Ons-O Grove, Sálvora-Corrubedo). Los SNPs *outliers* evidenciaron adaptación local (Fig. 3).

En RStudio se implementaron flujos de trabajo reproducibles, desde la importación de datos filtrados hasta la generación de DAPC y gráficos de estructura genética, facilitando una interpretación clara y visual de los resultados.

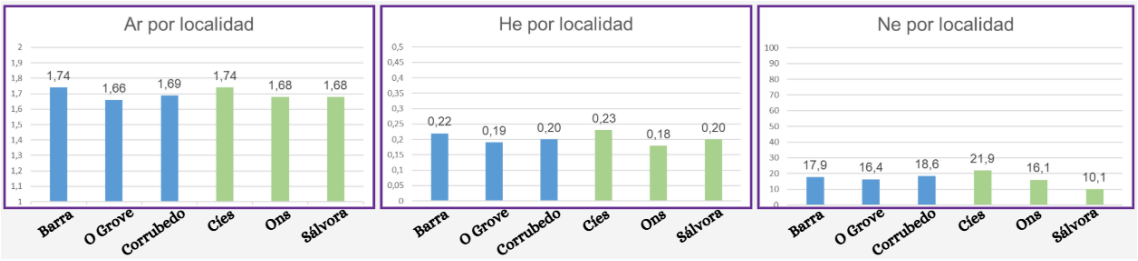



**Figura 2.** Riqueza alélica (Ar), heterocigosis esperada (He) y censo efectivo (Ne) obtenidas por localidad. En color verde las localidades continentales y en azul las insulares.

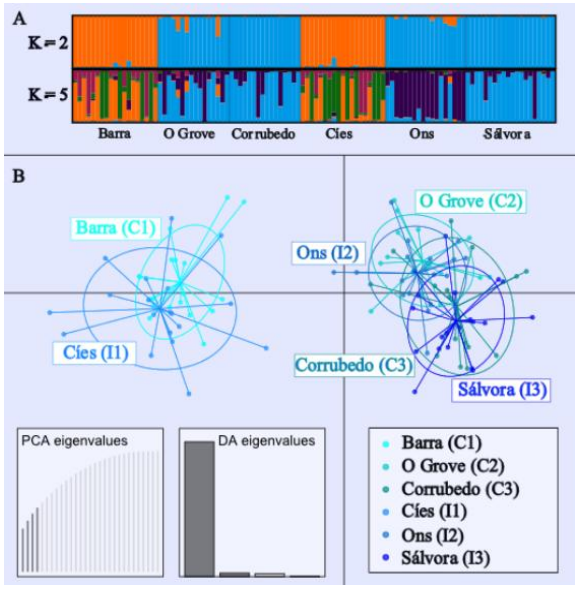



**Figura 3.** Análisis de estructura genética con el panel de 29 SNPs *outliers*. A) Agrupación genética representada por el análisis de STRUCTURE para K = 2 y K = 5. B). Análisis discriminante de los componentes principales (DAPC) representando el primer y segundo factor discriminante (DF) incluyendo el peso del análisis discriminante retenido (DA) para representar el 52,6% de la varianza.

## 5. CONCLUSIONES

- Con este trabajo se comprobó la utilidad de R y su catálogo de paquetes para los análisis de genómica de la conservación con recursos locales como de supercomputación (CESGA).
- Diversidad genética moderada en poblaciones insulares y continentales, pero con bajo tamaño efectivo.
- Estructuración local en tres pares isla-continente: Cíes-Barra, Ons-O Grove y Sálvora-Corrubedo.
- Evidencias de adaptación local en algunos SNPs. Recomendaciones: conservar in situ, recolectar germoplasma, seguimiento poblacional y aplicar este enfoque a otras especies/hábitats atlánticos.

**AGRADECIMIENTOS**

Beca de Colaboración

**Referencias bibliográficas**

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# ESTUDIO DE LA COMUNIDAD DE DIATOMEAS EN EL RÍO LOR: ANÁLISIS DE LA VARIABILIDAD Y RIQUEZA A NIVEL LOCAL

Olañeta D.[1], Gómez-Rodríguez C.[2] e Manel L.[1]

[1] Dpt. Biología Funcional (Ecología), Universidade de Santiago de Compostela
[2] CRETUS, Dpt. Biología Funcional (Ecología), Universidade de Santiago de Compostela

## RESUMEN

El seguimiento a medio y largo plazo de las comunidades biológicas es clave para detectar señales tempranas de posibles amenazas a la diversidad y funcionamiento de los ecosistemas. En el caso de los sistemas acuáticos, las diatomeas son utilizadas como bioindicadores del estado ecológico de los ríos en la Directiva Marco del Agua. Dada su importancia, es esencial identificar las localidades idóneas, y establecer un procedimiento óptimo, para la caracterización y seguimiento de las comunidades de diatomeas.

En el presente estudio se caracteriza la composición, estructura y variabilidad local de la comunidad de diatomeas en diferentes localidades del principal sistema fluvial de la Serra do Caurel, el río Lor, a su paso por Seoane do Courel (Lugo). El objetivo principal es identificar la localidad más diversa y representativa de este tramo del río, lo que la convertiría en el lugar de mayor interés para el seguimiento a largo plazo.

También se estudia si el esfuerzo de muestreo aceptado como estándar para estudios de diversidad específica en comunidades epilíticas de agua dulce (n = 500 valvas) es adecuado para la correcta caracterización de la comunidad. Se encuentra que dicha adecuación varía en función de los objetivos del estudio, poniendo en entredicho el actual método de empleo de estos organismos para el seguimiento ambiental.

Se resalta también cómo dicho incremento en el esfuerzo de muestreo tiene un impacto reducido a la hora de inferir el estado ecológico del entorno fluvial empleando índices basados en diatomeas. Esto no se debe a una falta de sensibilidad por parte del índice frente a cambios en el esfuerzo de muestreo, sino de una laguna de conocimiento en lo que respecta a los valores de indicatividad ecológica de las especies encontradas en este río gallego.

**Palabras y frases clave:** bioindicadores, diatomeas, diversidad alfa, diversidad beta, esfuerzo de muestreo

## 1. INTRODUCCIÓN

Las diatomeas constituyen un grupo de microalgas ampliamente utilizadas en la evaluación del estado ecológico de los ecosistemas acuáticos. La importancia de estos organismos como bioindicadores ha llevado al desarrollo de diversos índices bióticos basados en la abundancia relativa de especies ponderadas por su tolerancia

ecológica, cuyo uso está cada vez más extendido en programas de seguimiento ambiental, incluso formando parte de la normativa en varios países (Lobo et al., 2016; European Committee for Standardization, 2003), incluida España.

El uso efectivo de diatomeas como bioindicadores exige fiabilidad de los datos y, por tanto, la adecuación del esfuerzo de muestreo. Un esfuerzo de muestreo insuficiente puede dar lugar a modelos poco robustos e inferencias erróneas, mientras que un esfuerzo excesivo puede implicar un uso innecesario de recursos (Bennett et al., 2016). En el caso de las diatomeas, con estimaciones de entre 30.000 y 100.000 especies, muchas de las cuales son raras, se recomienda habitualmente contar entre 400 y 600 valvas por muestra (Bennett et al., 2016). Sin embargo, en las últimas décadas han surgido múltiples iniciativas que cuestionan la adecuación de dicho esfuerzo de muestreo, estudiando el efecto que el incremento o la reducción del número de valvas contados fuera del umbral recomendado tiene en la fiabilidad de los resultados (Bennett et al., 2016; Riato et al., 2024; Tyree et al., 2020).

En este trabajo se pone de manifiesto la excelente capacidad de estos organismos para la bioindicación del estado ecológico de los ecosistemas acuáticos, teniendo una amplia aplicabilidad a la conservación de estos hábitats en Galicia. Se aportan también evidencias empíricas al debate sobre la suficiencia del esfuerzo de muestreo estándar al mismo tiempo que se resaltan una serie de limitaciones metodológicas que el actual sistema de aplicación de las diatomeas presenta.

## 2. METODOLOGÍA

Se seleccionaron cuatro estaciones de muestreo en el tramo alto del río Lor, en las proximidades de Seoane do Courel. En cada muestra se contabilizaron valvas individuales de diatomeas hasta alcanzar un mínimo de 500 valvas por muestra, siguiendo los estándares habituales empleados en estudios de diversidad específica en comunidades epilíticas de agua dulce. Posteriormente, se amplió el esfuerzo de muestreo hasta alcanzar las 750 valvas por muestra, en incrementos sucesivos de 50 valvas. Este incremento permitió evaluar el efecto del tamaño muestral sobre los estimadores de riqueza y diversidad, así como mejorar la fiabilidad de los análisis cuantitativos.

Se caracterizó la diversidad alfa, es decir, el número y diversidad de especies, de cada localidad, esta última a través de índices (índice de diversidad de Shannon y de dominancia de Simpson) calculados con la función diversity() del paquete vegan (Oksanen et al., 2020) de R (R Core Team, 2024). También se comparó gráficamente la distribución de la abundancia en cada una de las localidades mediante la representación de SADs (Species Abundance Distributions) de las muestras completas.

La diversidad beta (disimilitud biótica) entre las muestras se calculó mediante el índice de Bray-Curtis, utilizando los datos de abundancia absoluta de especies. Para el cálculo se usó la función vegdist() del paquete vegan. Para visualizar el patrón de disimilitud, se realizó un escalamiento multidimensional no métrico (NMDS) en dos dimensiones mediante la función metaMDS() de vegan. Se representó el resultado con la función de R ggordiplots(), del paquete homónimo (Ramírez-Barahona, 2023). Para determinar cómo las diferencias en el número de valvas contadas en cada localidad afectaban a la fiabilidad del muestreo, se realizó una estimación de la riqueza aplicando el índice Chao1 a los datos de abundancia absoluta a través de la función estimateR() de vegan.

Finalmente, se analizó el efecto del tamaño muestral sobre la robustez del cálculo de los índices bióticos derivados de la composición de las comunidades, empleando índices basados en IPS (Specific Polluosensitivity Index), que son los incorporados en la Directiva Marco del Agua. Para ello se empleó la función diat_ips(), del paquete

diathor (Gelis et al., 2022) de R. Este cálculo se aplicó a los datos registrados tras el recuento de 500 y 750 valvas.

## 3. RESULTADOS

El análisis de la composición muestra que la localidad óptima para un seguimiento a largo plazo de la comunidad de diatomeas epilíticas del alto Lor es la etiquetada como LOR1 (42,6219728, -7,1713203), pues se trata de la localidad en la que se halló la mayor riqueza de especies y la mayor diversidad, tanto tras contar 500 como 750 valvas, pudiéndose establecer por tanto como la más representativa de una comunidad con un elevado valor ecológico. Además, también se trata de la localidad en la que se registró mayor densidad de individuos por superficie de biofilm, lo que facilita la toma de muestras para el seguimiento.

Este trabajo también muestra que el incremento del esfuerzo de muestreo no se corresponde con un incremento equivalente en el número de especies detectadas, tal y como se espera al ser una relación asintótica. El hecho de que la riqueza observada se incremente en menos del 20% al aumentar el esfuerzo de muestreo en un 50% sugiere que, aunque existe un amplio número de especies no detectadas (i.e. diferencia entre la riqueza estimada y la riqueza observada), podría no ser rentable incrementar el esfuerzo de muestreo en determinado tipo de estudios.

Los estudios centrados en la detección de patrones composicionales compartidos por múltiples muestras son los que potencialmente se verían más beneficiados de un incremento del esfuerzo de muestreo, ya que en este estudio se encuentra que permitió observar una tendencia de homogeneización de las muestras, reduciéndose las diferencias bióticas entre localidades al incrementar el número de valvas procesadas.

El índice de estado ecológico IPS mostró ser robusto frente al incremento en el esfuerzo de muestreo, no estando justificado, en nuestro estudio, contar más de 500 valvas, pues no supondría una mejora notable en la estimación del índice. Sin embargo, esto se explica, no porque el índice no sea sensible a especies raras, sino por una limitación metodológica que nace de lagunas en las bases de datos empleadas en el cálculo.

En lo que respecta a los valores obtenidos del IPS, bajo los criterios utilizados en la implementación de la Directiva Marco del Agua, todas las localidades estudiadas en este trabajo pueden clasificarse como buen estado ecológico

## 4. CONCLUSIONES

Tras la identificación y cuantificación de la comunidad de diatomeas epilíticas en diferentes localidades próximas del río Lor a su paso por Seoane do Courel, y la caracterización de su estructura y variabilidad entre las localidades de muestreo, se establece la localidad LOR1 (42,6219728, -7,1713203) como la óptima para el seguimiento a largo plazo de la evolución de dicha comunidad.

Se establece que la adecuación del esfuerzo de muestreo estándar, o la necesidad de reducirlo optimizando la relación coste/beneficio, radica en el objetivo del estudio. Para aquellos análisis que dependan de una completa caracterización de la composición florística de la comunidad para su robustez se precisará de >600 valvas contadas, siempre que la comunidad sea sana y diversa. Por otro lado, cuando una elevada precisión taxonómica no sea necesaria para los objetivos del estudio, recuentos de menos de 400 valvas pueden ser suficientes.

Al analizar en qué medida el esfuerzo de muestreo afectó a la robustez de los índices, se puso de manifiesto una limitación metodológica adicional: el poco conocimiento del valor indicativo de las especies raras de cara a su aplicación en índices bióticos, por lo que también sugerimos el ahondamiento en esta cuestión en futuros trabajos.

Finalmente, en base a los resultados de los índices IPS se afirma que el agua del río Lor en su tramo alto encaja dentro de la categoría de buen estado ecológico, de acuerdo con los umbrales establecidos por la normativa vigente.

En definitiva: LOR1 es la mejor localidad para un seguimiento a largo plazo. Contar más de 500 valvas no mejora significativamente los índices, por lo que no estaría justificado el gasto de tiempo y recursos, pero puede ser recomendable para ciertos estudios. El río Lor presenta buen estado ecológico.

**Referencias**


Bennett, J. R., Sisson, D. R., Smol, J. P., Cumming, B. F., Possingham, H. P. y Buckley, Y. M. (2014): "Optimizing taxonomic resolution and sampling effort to design cost-effective ecological models for environmental assessment", Journal of Applied Ecology, 51(6), pp. 1722–1732.

European Committee for Standardization (2003): Water quality – Guidance standard for the routine sampling and pre-treatment of benthic diatoms from rivers. European Standard EN 13946:2003, Brussels, CEN.

Gelis, M. N., Sathicq, M. B., Jupke, J. y Cochero, J. (2022): "DiaThor: R package for computing diatom metrics and biotic indices", Ecological Modelling, 465, 109859.

Lobo, E. A., Heinrich, C. G., Schuch, M., Wetzel, C. E. y Ector, L. (2016): "Diatoms as bioindicators in rivers", en B. R. Singh, ed., River algae, Cham, Springer, pp. 245–271.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D. et al. (2020): vegan: Community Ecology Package (R package version 2.5-7).

Ramírez-Barahona, S. (2023): ggordiplots: Plots for ordination results using ggplot2 (R package version 0.4.1).

R Core Team (2024): R: A language and environment for statistical computing, Vienna, R Foundation for Statistical Computing.

Riato, L., Stoddard, J. L., Herlihy, A. T. y Blocksom, K. A. (2024): "Reduced count size can provide a robust and more efficient diatom assessment of environmental conditions", Journal of Applied Ecology, 61(9), pp. 2308–2320.

Tyree, M. A., Carlisle, D. M. y Spaulding, S. A. (2020): "Diatom enumeration method influences biological assessments of southeastern USA streams", Freshwater Science, 39(1), pp. 183–195.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# O MANEXO DE IMAXES TERMOGRÁFICAS CON R: IDENTIFICANDO O PERÍMETRO DUN INCENDIO FORESTAL

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

[1]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga).
[2]Grupo Modestya, Departamento de Estatística, Análise Matemático e Optimización, Universidade de Santiago de Compostela.

## RESUMO

Este traballo presenta un algoritmo desenvolto integramente con R, que permite identificar o perímetro dun incendio forestal en tempo real a partir das imaxes termográficas tomadas polas aeronaves de extinción. Presentarase a metodoloxía desenvolta, as librerías e funcións de R utilizadas, e algunhas imaxes de perímetros de incendios reais construídas mediante o algoritmo descrito.

**Palabras e frases chave**: Imaxes termográficas, incendios forestais, librería terra, librería dplyr.

## 1. INTRODUCIÓN

Coñecer en tempo real o perímetro dun incendio forestal é de gran interese para a xestión das tarefas de extinción. A localización e a forma exacta do perímetro pode conseguirse mediante fotografías termográficas totais ou parciais do incendio, que poden ser realizadas coa axuda de drons ou das aeronaves de extinción. O obxectivo do algoritmo que se presenta é obter información do perímetro dun incendio en tempo real de xeito automático, evitando así un aumento da carga de traballo para o persoal que xestiona a actuación no incendio.
Para o desenvolvemento deste algoritmo, cóntase con imaxes sacadas por aeronaves de extinción durante o seu traballo no incendio forestal. As imaxes que se obteñen son imaxes parciais do incendio, nas que cada píxel ten un valor de temperatura asociada.
O procesado das imaxes termográficas de incendios forestais presenta algunhas dificultades. Por exemplo, non se poden establecer valores de temperaturas fixos a partir dos cales un píxel da imaxe termográfica pertence ou non ó incendio, xa que hai incidencias que poden alterar os valores de temperatura da imaxe, como pode ser un problema de refrixeración do sensor termográfico. Tamén se debe ter en conta que non todos os incendios acadan os mesmos valores de temperatura, os cales poden depender, entre outras cousas, da velocidade do incendio ou do tipo de combustible que se está a queimar.

## 2. SOLUCIÓN APORTADA

O obxectivo principal do traballo proposto é identificar o perímetro dun incendio forestal en tempo real e de xeito automático a partir de imaxes termográficas parciais do mesmo. Para isto, implementouse en R un algoritmo que recibe como dato de entrada a ruta e o nome dunha carpeta onde se atopan todas as imaxes termográficas do incendio. Estas imaxes deben ser en formato **tif**.
Para cada unha das imaxes que se atopan nesa carpeta, créase un obxecto ráster empregando a función *rast* da librería *terra*. A continuación, descártanse os píxeles con valores negativos, xa que se consideran errores do sensor, e constrúese un obxecto *dataframe* coas coordenadas e a temperatura correspondente. Despois aplícase un algoritmo **K-Means** mediante a función *kmeans* da librería

*stats*, que permite agrupar os píxeles do incendio (representados polas súas coordenadas) en 3 grupos en función da súa temperatura (representando os píxeles non queimados, os queimados, e nos que o incendio está activo no momento de obter a imaxe). Isto permite establecer o valor a partir do cal se considera que un píxel se corresponde ou non cun incendio, sen ter que establecer valores absolutos de temperatura.
Selecciónanse todos os píxeles que teñen un valor de temperatura maior ou igual ó establecido no punto anterior e convértese o *dataframe* correspondente a un obxecto espacial e calcúlase a súa envolvente convexa, empregando funcións da librería *sf*. Deste xeito, tense a xeometría do incendio na imaxe analizada.
Este procedemento repítese para cada unha das imaxes, unindo os perímetros que se van calculando, para ó final facer a envolvente convexa do conxunto, que representará o perímetro do incendio. Ademais, únense todas as imaxes que se usan como entrada creando un mosaico das mesmas, o que permite ter unha visualización máis completa do incendio activo.
O algoritmo devolve como saída unha imaxe do perímetro calculado e tamén o mosaico de todas as imaxes que se usan como entrada. Un exemplo de saída do algoritmo móstrase na Figura 1.

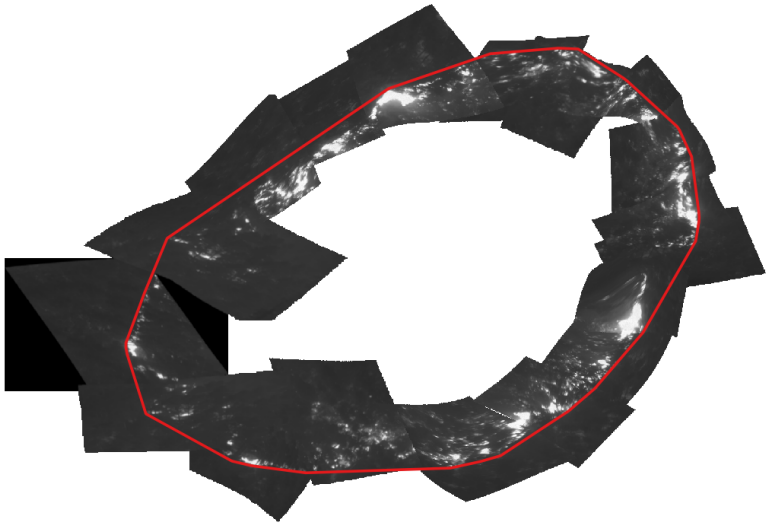


Figura 1: Exemplo de saída de execución do algoritmo desenvolto: mosaico das imaxes termográficas utilizadas como dato de entrada e, en vermello, liña que representa o perímetro do incendio.

O algoritmo está empaquetado nun contedor Docker, o que permite que poda ser distribuído e executado de forma sinxela en calquera sistema operativo.

# Referencias


[1] Hijmans, R. (2024). terra: Spatial Data Analysis. R package version 1.7-71, https://CRAN.R-project.org/package=terra

[2] Ikotun A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

[3] Pebesma, E., Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall. https://doi.org/10.1201/9780429459016

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

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# O USO DE R EN SAÚDE PÚBLICA

Miguel Ángel Rodríguez Muiños, Gael Naveira Barbeito

Dirección Xeral de Saúde Pública, Consellería de Sanidade, Xunta de Galicia

## RESUMO

O relatorio versará sobre a utilización de R no ámbito da Saúde Pública. Falaremos dos visores de datos programados en Shiny como poden ser:

- Catálogo de Publicacións
- Resultados Analíticos das Augas de Baño 2008-2024
- Sistema de Información sobre Mortalidade por Cancro de Galicia
- Mortalidade por causas seleccionadas
- Radiant on-line

Tamén falaremos sobre o Proxecto BioStatFLOSS, que consiste nun ámbito unificado e homoxéneo de execución, baixo o sistema operativo Microsoft Windows, dunha recompilación de programas especificamente deseñados para a realización de estudos epidemiolóxicos, bioestatísticos e de saúde en xeral. O proxecto consiste nunha revisión e preparación do software relevante seleccionado e o desenvolvemento dun ámbito de execución común dende o que poder lanzar devanditos programas. O obxectivo, dotar dunha infraestrutura FLOSS (Free/Libre Open Source Software) e portable (que non necesite instalación) aos usuarios deste tipo de programas facilitando a migración do software privativo a solucións de software libre.

**Palabras e frases chave:** R, Shiny, BioStatFLOSS, Saúde, Administración

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# Estimación robusta de rexións de referencia condicionais. Aplicación na investigación da diabetes

Sara Rodríguez-Pastoriza[1], Javier Roca-Pardiñas[1,2] e Óscar Lado-Baleato[3,4]

[1]Department of Statistics and Operations Research, University of Vigo, 36310 Vigo, Galicia, Spain.
[2]Galician centre for mathematical research and technology, CITMAGA, Galicia, Spain.
[3]Research Methods Group (RESMET), Health Research Institute of Santiago de Compostela (IDIS), Galicia, Spain.
[4]ISCIII Support Platforms for Clinical Research, Health Research Institute of Santiago de Compostela (IDIS), Galicia, Spain.

## RESUMO

O diagnóstico de varias enfermidades básase en dous ou máis marcadores diagnósticos continuos. Os resultados destos marcadores soen estar correlacionados, e a súa distribución conxunta pode variar con covariables clínicas (por exemplo, a idade), independientemente do estado da enfermidade. Nestos casos, as rexións de referencia multivariantes, que caracterizan ónde se espera que se sitúen o 95 % dos resultados das persoas sás, ofrecen unha regra de interpretación para diagnosticar estas enfermidades. A estimación de rexións de referencia multivariantes condicionais, que se adaptan aos valores das covariables, é crucial para o diagnóstico personalizado. Sen embargo, segue sendo un reto metodolóxico, especialmente en presenza de efectos non lineais das covariables e valores atípicos na variable de resposta.

No presente traballo propoñemos un novo modelo de regresión bivariada non paramétrica para estimar rexións de referencia multivariantes condicionais. A nosa proposta baséase na regresión cuantílica aditiva flexible e non asume ningunha distribución paramétrica específica para os datos. Non obstante, asumimos a convexidade da distribución multivariante da variable resposta, unha suposición razoable na maioría das aplicacións clínicas. O método primeiro estima o centroide axustado por covariables da distribución (mediana bivariada) e, a continuación, constrúe cuantís direccionais para definir unha rexión de predición que proporciona unha cobertura uniforme en todas as direccións.

As principais vantaxes do noso método inclúen: (1) robustez fronte a valores atípicos, o que mellora a estabilidade en aplicacións con datos do mundo real; (2) flexibilidade no modelado de efectos covariables non lineais, utilizando funcións spline penalizadas e cíclicas; (3) estimación libre de distribución, o que o fai aplicable tanto a datos gaussianos como non gaussianos; e (4) comportamento de cola igual, o que garante que a mesma porcentaxe de datos quede fora da rexión en todas as direccións.

Os estudos de simulación demostran a precisión, solidez e fiabilidade do método proposto tanto en distribucións gaussianas como non gaussianas, con e sen contaminación por valores atípicos. Validamos aínda máis o noso enfoque empregando un conxunto de datos do mundo real que inclúe dous biomarcadores glucémicos comúnmente utilizados no diagnóstico da diabetes (glucosa plasmática en xaxún e porcentaxe de hemoglobina glicosilada). O modelo captura eficazmente a influencia da idade na distribución conxunta destos marcadores e proporciona rexións de referencia individualizadas e clínicamente significativas. Este marco ofrece unha ferramenta práctica e robusta para mellorar a interpretación das probas diagnósticas correlacionadas de forma personalizada.

Finalmente, como liña complementaria, tamén exploramos un segundo enfoque orientado a detectar de maneira directa a influencia dunha covariable sobre a resposta

bivariada en función do cuantil, o que permite identificar efectos que só aparecen en rexións específicas da distribución e non se manifestan a nivel global.

Todo o desenvolvemento metodolóxico, simulacións e aplicacións prácticas realizáronse integramente en R, empregando os paquetes qgam, mgcv e refreg, o que garante a reproducibilidade dos resultados e a dispoñibilidade aberta da metodoloxía.

*XII Xornada de Usuarios de R en Galicia*

*Santiago de Compostela, 16 de outubro do 2025*

# Análisis funcional y modelado predictivo de energía fotovoltaica con R: un estudio en la UDC

Marina Sánchez Villaverde, Manuel Oviedo de la Fuente e Carlos José Escudero Cascón

[1] Universidade da Coruña
[2] Universidade da Coruña y CITIC
[3] Universidade da Coruña y CITIC

**RESUMO**

Este Trabajo se centra en el análisis y modelado de la energía solar fotovoltaica a partir de datos recogidos en las instalaciones de la Universidade da Coruña, concretamente en los edificios de la Facultad de Derecho y del Departamento de la Escuela Técnica Superior de Arquitectura.

Dado que estas instalaciones son recientes, el trabajo constituye el primer estudio sobre su funcionamiento, proporcionando información clave sobre su eficiencia y sirviendo como base para futuras iniciativas sostenibles dentro de la universidad y en otros contextos.

La motivación principal del estudio es promover el uso de energías renovables, menos dañinas para el medio ambiente, y explorar el potencial real de la energía solar en Galicia, donde las condiciones meteorológicas son variables.

En la primera fase, se realiza un análisis detallado de los datos, incluyendo detección de valores atípicos, suavizado de curvas, estudio de correlaciones, reducción de dimensionalidad mediante PCA y agrupación de patrones mediante clustering.

En la segunda fase, se desarrollan modelos predictivos (lineales y aditivos) para estimar la producción de energía y evaluar cómo los factores climáticos influyen en su comportamiento.

Los resultados obtenidos permiten evaluar la eficacia de las instalaciones y generar herramientas predictivas que faciliten la planificación y gestión energética basada en datos reales, adaptados a las condiciones locales, con un enfoque práctico y aplicable a otras instalaciones universitarias.

**Palabras e frases chave:** Inclúa aquí as palabras e frases chave. Máximo de seis.

Energía fotovoltaica, energía renovable, análisis de datos funcionales, modelado predictivo, datos meteorológicos.

## 1. INTRODUCIÓN

En la actualidad, la contaminación y el cambio climático representan retos urgentes que requieren una transformación del modelo energético. En este contexto, las energías renovables ganan protagonismo, aprovechando recursos naturales como el sol y el viento para generar electricidad con menor impacto ambiental.

España ha incrementado notablemente el uso de energías limpias: entre 2004 y 2022, el porcentaje de energía renovable sobre el consumo total pasó de un 8,35 % a más del 22 %. En particular, la energía solar fotovoltaica ha experimentado un crecimiento destacado desde 2019, tanto en potencia instalada como en energía generada, impulsado por políticas de apoyo y una mayor concienciación ambiental.

El consumo de energía fósil ha provocado problemas como el calentamiento global, contaminación y generación excesiva de residuos. La sociedad y organismos internacionales fomentan así un consumo responsable y sostenible, promoviendo la energía verde obtenida de fuentes naturales inagotables y con bajo impacto ambiental.

Sin embargo, muchas energías renovables dependen de las condiciones meteorológicas, lo que plantea desafíos para garantizar un suministro estable. Por ello, comprender cómo influyen las variables climáticas en la producción solar es esencial para planificar y gestionar la demanda energética de manera eficiente.

Este trabajo se centra en analizar la producción de energía solar fotovoltaica y su relación con variables meteorológicas y consumo, utilizando técnicas de análisis de datos y modelos predictivos para apoyar la gestión eficiente de recursos renovables y contribuir a un sistema eléctrico más sostenible.

## 2. SÍNTESIS

El trabajo analizó la producción y consumo energético en la Universidade da Coruña a partir de datos de 10 minutos transformados en curvas funcionales con R. Tras la exploración inicial y la depuración de anomalías, se evaluaron distintos métodos de suavizado, aunque se decidió conservar los datos originales para no perder información relevante en picos y comportamientos extremos. Con la correlación de distancias se detectaron dependencias no lineales entre consumo, producción y factores meteorológicos o de calendario, y mediante técnicas de profundidad y derivadas se identificaron días atípicos asociados a fenómenos meteorológicos o usos no habituales de los edificios.

La reducción de dimensionalidad mediante FPCA explicó más del 80 % de la variabilidad y permitió representar los patrones principales, que se integraron en un análisis de clustering junto con variables meteorológicas y de calendario, revelando diferencias claras en la respuesta de cada edificio. En la fase de modelización se compararon regresiones lineales, modelos aditivos funcionales y Boosting, combinando variables funcionales, meteorológicas y de calendario. La validación con rolling windows y cross-validation mostró que los modelos aditivos funcionales capturaban mejor las relaciones no lineales, mientras que los métodos de Boosting resultaron eficaces para predecir la producción fotovoltaica, especialmente en la Escuela de Arquitectura.

## 3. CONCLUSIONES

Este trabajo ha analizado en detalle la producción y consumo de energía en la Facultad de Derecho y la Escuela de Arquitectura de la Universidade da Coruña, combinando técnicas de análisis funcional, modelos estadísticos avanzados y validación empírica.

Se identificaron patrones claros de consumo, diferenciando entre días laborables y no laborables: Derecho muestra un descenso significativo del consumo los fines de semana, mientras que Arquitectura mantiene un patrón más estable. Además, la

producción solar se ve influida por el sistema de autoconsumo y por factores meteorológicos, destacando la temperatura como variable relevante en Arquitectura.

La detección de valores atípicos permitió identificar días con comportamientos extremos, relacionados con eventos meteorológicos o usos no registrados de los edificios. El uso de Análisis de Componentes Principales Funcional (FPCA) y clustering funcional mostró diferencias estructurales entre edificios y confirmó cómo responden de manera diferenciada a factores externos según su arquitectura y uso.

En la fase de modelización, los modelos aditivos funcionales (GAM) demostraron ser especialmente útiles para capturar relaciones no lineales y efectos suaves. Los mejores modelos, que combinaron variables meteorológicas, funcionales y de calendario, explicaron hasta el 85 % de la variabilidad en la producción y el 92 % en la desviación

de los datos. Los modelos de boosting también mostraron un buen desempeño, especialmente para predecir la producción fotovoltaica en la Escuela de Arquitectura.

En conjunto, estos resultados permiten comprender mejor el comportamiento energético de los edificios y desarrollar herramientas predictivas que faciliten la planificación y gestión eficiente de recursos renovables, adaptadas a las condiciones locales y a los patrones de uso específicos de cada instalación.

**Referencias**

[1] Febrero-Bande, M., & De La Fuente, M. O. (2012). Statistical computing in functional dataanalysis: The R package fda. usc. Journal of statistical Software, 51, 1-28.

[2] "Funcionamiento de las placas fotovoltaicas," Iberdrola, 2024. [En línea]. Disponible en:https://www.iberdrola.com/innovacion/como-funcionan-placas-solares fotovoltaicas

[3] Edelmann, D., Móri, T. F., & Székely, G. J. (2021). On relationships between the Pearsonand the distance correlation coefficients. Statistics & probability letters, 169, 108960.

[4] Ferraty, F., Laksaci, A., & Vieu, P. (2006). Estimating some characteristics of theconditional distribution in nonparametric functional models. Statistical Inference forStochastic Processes, 9, 47-76.

[5] Ramsay, J. O., & Silverman, B. W. (2005). Principal components analysis for functionaldata. Functional data analysis, 147-172.

[6] Fernández-Casal, R., Costa, J., & Oviedo de la Fuente, M. (2024). Métodos predictivos deaprendizaje estadístico.

[7] Febrero-Bande, M., González-Manteiga, W., & Oviedo De La Fuente, M. (2019). Variableselection in functional additive regression models. Computational Statistics, 34, 469-487.

*XII Xornada de Usuarios de R en Galicia*
*Santiago de Compostela, 16 de outubro do 2025*

# O uso da cor na difusión de datos estatísticos: unha cuestión nada trivial. Quarto e Typst: alternativas a RMarkdown e LaTeX

Jordán Soubrier Benavides[1]

[1]Instituto Galego de Estatística (IGE)

## RESUMO

A selección dunha paleta de cor axeitada é unha das decisións máis críticas e, desafortunadamente, con frecuencia infravaloradas no contexto da visualización gráfica de datos. Unha paleta inadecuada pode conducir tanto a distorsión na percepción dos datos como a problemas serios de accesibilidade e comunicación científica. Nesta exposición explóranse as razóns polas que a escolla de paletas vai moito máis alá do criterio estético, abordando as diferenzas esenciais entre os espazos de cor (RGB, HCL, Lab...) e os requisitos que debe cumprir unha paleta adecuada segundo os estándares científicos actuais. Aínda que o suxeito desta exposición pode resultar sinxelo en apariencia, calquera que se adentre no estudo da materia veráse fácilmente abrumado pola complexidade que reviste a cuestión.

Preténdese achegar aos usuarios de R unha explicación básica desta problemática, suxerindo ferramentas adecuadas e de uso sinxelo.

A característica principal que debe ter unha paleta axeitada é a uniformidade perceptual, isto é, que a magnitude da diferencia percibida entre as diferentes cores se corresponda cas diferencias existentes nos datos representados.

A segunda característica é a inclusividade cara ás persoas con deficiencias na percepción da cor (daltonismo). A uniformidade perceptual debe conservarse nos espazos de cor bidimensionais asociados ao dicromatismo: deuteranopía, propanopia e tritanopia, así como na monocromía. Isto ten a vantaxe adicional de dar –con frecuencia, non sempre– resultados axeitados á impresión en branco e negro.

Ademais disto, o uso que se pretenda dar á paleta tamén condiciona a súa concepción e deseño. Segundo o tipo de datos que se vaian representar, as paletas poden ser continuas, discretas ou categóricas (tipos) e, nunha dimensión adicional (clases), secuenciais, diverxentes –para amosar desviacións respecto dun valor de referencia–, multisecuenciais –para a comparación simultánea de varias categorías ou grupos– e mesmo cíclicas –axeitadas para datos periódicos, como a hora do día, a dirección de fluxos ou fases dun ciclo–.

Aínda que no día a día o usuario de R adoita traballar con códigos hexadecimais, este espazo é unha discretización do espazo RGB, un modelo de cor aditivo adaptado ás pantallas. É moi habitual empregar a función `colorRamp()` para crear gradientes entre dúas cores. Sen parámetros adicionais, o resultado é unha interpolación lineal no espazo RGB –que non é perceptualmente uniforme–, e polo tanto inadecuado.

A análise e xeración de mapas de cores perceptualmente uniformes debe por tanto realizarse en espazos de cor que teñan a percepción humana como referencia. Sen profundizar en teoría e formulación, exploraremos dous espazos de cor enfocados na percepción humana e a uniformidade perceptual, O HCL (Hue-Chroma-Luminance) e o CIE Lab. Presentaremos dous paquetes de R que proporcionan paletas perceptualmente uniformes nestes espazos: `colorspace` e `scico`.

Examínase o paquete `colorspace`, que ademáis de incluir un conxunto de paletas elaboradas no espazo HCL, contén unha excelente serie de ferramentas para a diagnose

e xeración de paletas perceptualmente uniformes, que complementa cunha documentación moi clarificadora sobre os espazos de cor, a uniformidade perceptual e os tipos e clases de paletas, e proponse este paquete como unha das alternativas máis potentes na escolla de paletas para a representación de datos.

A continuación revisamos o paquete `scico` e o traballo no que se fundamenta, *Scientific Colour Maps*, un conxunto completo –para todos os tipos e clases– de paletas elaboradas no espazo Lab que, ademais de cumprir cos dous requisitos fundamentais, engade outros adicionais: ordenación perceptual, lexibilidade na impresión en branco e negro, citabilidade e reproducibilidade. Este conxunto de paletas constitúe probablemente a opción máis axeitada e rigorosa actualmente.

Comparamos os resultados destas paletas co de paletas inadecuadas, facendo evidente a necesidade de seguir estes criterios na presentación visual de datos estatísticos ou científicos.

Finalmente, fronte a necesidade de incluir nas paletas, por exemplo, unha cor corporativa, o autor presenta as súas indagacións e código experimental para abordar a elaboración de paletas personalizadas respectando os criterios establecidos, poñendo de relevo a dificultade desta empresa.

Como apéndice, e moi brevemente, danse a coñecer dúas pezas de software libre moi útiles para os usuarios de R, especialmente aqueles que habitualmente empregan o formato `RMarkdown` para as súas publicacións ou documentación:

`Quarto` é unha evolución do concepto de reproducibilidade e publicación científica iniciado por RMarkdown. É multilinguaxe –soporta `Python`, `Julia`, `Javascript` ou `R` indistintamente–, as referencias a figuras, táboas, seccións, ecuacións e código son máis flexibles e sinxelas, ten maiores capacidades de personalización de estilos e formatos, compatibilidade con diversos IDEs, e renderizado directo a múltiples formatos.

`PDF` a través de `Typst` é un dos formatos de saída que soporta Quarto. `Typst` é unha alternativa moderna a LaTeX cunha sintaxe básica moi similar a `Markdown` e unha linguaxe de programación sinxela e moi flexible, cunha curva de aprendizaxe accesible, e unha compilación moi rápida.

**Palabras e frases chave**: data visualization, perceptually uniform palettes, colorspaces, scico, Quarto, typst

# Referencias

[1] Crameri, F. (2018). Scientific colour maps. Recuperado de `https://www.fabiocrameri.ch/colourmaps/`

[2] Crameri, F., Shephard, G.E., Heron, P.J. (2020). The misuse of colour in science communication. *Nature Communications*, 11, 5444. `https://doi.org/10.1038/s41467-020-19160-7`

[3] Ihaka, R., Murrell, P., Hornik, K., Fisher, J.C., Stauffer, R., Wilke, C.O., McWhite, C.D., Zeileis, A. (s.f.). *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes.* Recuperado de `https://colorspace.r-forge.r-project.org/`

[4] Pedersen, T.L., Grund, M., Campos, F., Clark, M. (s.f.). *scico: Palettes for R based on the Scientific Colour-Maps.* Recuperado de `https://github.com/thomasp85/scico`

[5] Kovesi, P. (2015). Good Colour Maps: How to Design Them. Recuperado de `https://colorcet.com/`

**AUTORES**

# XII XORNADA DE USUARIOS DE R EN GALICIA

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

> ORGANIZA

Melisa

> PATROCINAN

amtega
Axencia para a Modernización Tecnolóxica

XUNTA DE GALICIA

ISBN X XXXXXX XXXXXX