# XV Congreso Galego de Estatística e Investigación de Operacións


#SGaPEIO

Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DEPARTAMENTO DE ESTATÍSTICA, ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

# XV Congreso Galego de Estatística e Investigación de Operacións

# Libro de actas


**Santiago de Compostela, 4, 5 e 6 de Novembro de 2021**

XV Congreso Galego de Estatística
e Investigación de Operacións.
Libro de actas

Elaborado por:
Comité Organizador

I.S.B.N.: 978-84-09-35306-4

# Comité científico

- Gustavo Bergantiños Cid (Universidade de Vigo)
- Carmen María Cadarso Suárez (Universidade de Santiago de Compostela)
- Balbina Virginia Casas Méndez (Universidade de Santiago de Compostela)
- Ricardo Cao Abad (Universidade de A Coruña)
- Rosa María Crujeiras Casais (Universidade de Santiago de Compostela)
- Jacobo de Uña Álvarez (Universidade de Vigo)
- Isabel del Río Viqueira (Instituto Galego de Estatística)
- Manuel Febrero Bande (Universidade de Santiago de Compostela)
- Gloria Fiestras Janeiro (Universidade de Vigo)
- Ignacio García Jurado (Universidade de A Coruña)
- Julio González Díaz (Universidade de Santiago de Compostela)
- Wenceslao González Manteiga (Universidade de Santiago de Compostela)
- María José Lombardía Cortiña (Universidade de A Coruña)
- Covadonga Rodríguez Moldes Rey (Sociedade Galega de Estatística e Investigación de Operacións)
- Inês Sousa (Universidade do Minho)

# Comité organizador

- César Andrés Sánchez Sellero (Presidente)
- María José Ginzo Villamayor (Secretaria)
- Jose Ameijeiras Alonso
- María Isabel Borrajo García
- Mercedes Conde Amboage
- Milagros Diéguez Taboada
- Pedro Faraldo Roca
- María Esther López Vizcaíno
- María Martín Vila
- Alejandro Saavedra Nieves
- Paula Saavedra Nieves

# Presentación

Por petición da Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO), o Departamento de Estatística, Análise Matemática e Optimización da Universidade de Santiago de Compostela organiza o *XV Congreso Galego de Estatística e Investigación de Operacións*, que se celebra na Facultade de Matemáticas entre os días 4 e 6 de Novembro de 2021.

Este documento recolle os resumos das cinco conferencias plenarias, as tres contribucións da sesión de Biometría organizada conxuntamente pola SGAPEIO e a Sociedade Portuguesa de Estatística e as cincuenta e tres contribucións presentadas no congreso (corenta e sete en forma de comunicación oral e seis en formato póster). Agradecémoslles ás autoras e aos autores destes traballos as súas contribucións.

O Comité Científico
O Comité Organizador

# Contidos

**Venres, 5 de Novembro**

**16:00 - 17:30**



**Sábado, 6 de Novembro**

**9:30 - 10:45**

# Conferencias plenarias

# Una descripción general de la detección de atípicos en series temporales con avances recientes en grandes dimensiones

**Pedro Galeano San Miguel (Universidad Carlos III de Madrid)**

El objetivo de esta conferencia plenaria es el de realizar una descripción genérica de la detección de datos atípicos en series temporales univariantes y multivariantes. Además, se presentan algunas contribuciones muy recientes en el caso en el que la detección de atípicos se realice en un conjunto grande de series temporales. Estas contribuciones nuevas resultan del trabajo del ponente con los profesores Daniel Peña de la Universidad Carlos III de Madrid y Ruey S. Tsay de la Universidad de Chicago.

Habitualmente, los métodos y modelos estadísticos más comúnmente utilizados se basan en una serie de supuestos: Gaussianidad en muchos entornos, linealidad en modelos de regresión, llegadas exponenciales en la teoría de las colas, modelos ARIMA en series temporales, etc. A veces, los supuestos del método y/o modelo describen la mayoría de las observaciones, pero hay un número muy pequeño de ellas que tienen diferentes patrones. En estadística, se dice que estas observaciones son atípicas. En otras áreas, estas observaciones reciben diferentes nombres incluyendo los de observaciones aberrantes, anormales, anomalas, contaminantes, defectuosas, discordantes, excepcionales, nocivas, y un largo etcétera.

¿Por qué tratar con valores atípicos? En primer lugar, existen múltiples peligros resultantes de ignorar la presencia de estos datos. Por ejemplo, pueden introducir sesgos en las estimaciones de los parámetros del modelo, influir en la potencia de tests basados en tales estimaciones, aumentar la amplitud de los intervalos de confianza y predicción, o influir en predicciones, entre otros aspectos. Además, los valores atípicos suelen proporcionar información en una amplia variedad de aplicaciones. Por ejemplo, en movimientos de tarjetas de crédito, los valores atípicos pueden indicar usos fraudulentos de estas tarjetas, en redes de computadoras, los valores atípicos pueden indicar que una computadora ha sido pirateada, en registros médicos, los valores atípicos pueden indicar el empeoramiento de un paciente enfermo o la presencia de una enfermedad en una persona supuestamente sana, en apuestas deportivas, los valores atípicos pueden indicar amaño de partidos.

¿Cómo lidiar con los valores atípicos? En primer lugar, se puede utilizar métodos robustos. Estos métodos son resistentes a la presencia de valores atípicos o desviaciones de los supuestos del modelo. La ventaja principal de estos métodos es que proporcionan estimaciones fiables de los parámetros y predicciones más precisas. La desventaja principal es que se produce una perdida de eficiencia si no hay datos atípicos y una perdida potencial de la información que los atípicos pueden proporcionar. En segundo lugar, se

pueden utilizar procedimientos de detección de datos atípicos. Estos métodos detectan la presencia de los datos atípicos y toman acciones adecuadas para el análisis. Las ventajas principales de estos métodos es que se obtiene la información sustancial proporcionada por los datos atípicos, mantienen la complejidad del modelo baja y se pueden utilizar en combinación con los métodos robustos. La desventaja principal es que pueden producir falsos atípicos.

En series temporales, podemos encontrar datos atípicos que resultan del efecto de diferentes eventos en los valores de la serie temporal. Por ejemplo, una huelga en series temporales de producción, un evento de clima extremo en índices de contaminación del aire, un cambio drástico en el precio de un activo, etc. Tales eventos pueden producir datos atípicos aislados, secuencias de datos atípicos consecutivos o cambios de nivel, tendencia o varianza, entre otros aspectos. Además, detectar atípicos en series temporales en complejo debido a que las observaciones están autocorreladas, y los efectos sobre las series temporales dependen de las características de las series. Como consecuencia existen procedimientos de detección de observaciones atípicas (en el amplio sentido de la palabra) para modelos ARIMA, GARCH, SV, TAR e INGARCH, entre muchos otros.

La conferencia se divide en tres partes:

1. En primer lugar, se presentan los dos procedimientos más utilizados para la detección de datos atípicos en series temporales univariantes que siguen modelos ARIMA. Estos son los procedimientos propuestos por Chen y Liu (1993) y los procedimientos propuestos por Campos, Hendry y Johansen (2008) y Castle et el (2015). Además, se presenta un procedimiento muy reciente propuesto por Galeano, Peña y Tsay (2022a) basado en el uso de métodos de selección de variables para modelos de regresión lineal, que tiene un comportamiento mejor que los anteriores.

2. En segundo lugar, se presentan los dos procedimientos más utilizados para la detección de datos atípicos en series temporales multivariantes que siguen modelos VARMA. Estos son los procedimientos propuestos por Tsay, Peña y Pankratz (2000) y Galeano, Peña y Tsay (2006). Se comentan sus propiedades y sus problemas cuando el número de series es muy elevado.

3. En tercer lugar, se presenta un procedimiento recientemente propuesto por Galeano, Peña y Tsay (2022b) para la detección de atípicos en series temporales de alta dimensión. El procedimiento está basado en el uso de modelos factoriales dinámicos que permiten distinguir datos atípicos que afectan a muchas de las series, con datos atípicos que afectan a un gran número de series.

## REFERENCIAS

[1] Campos, C., Hendry, D. y Johansen, S. (2008) Automatic selection of indicators in a fully saturated regression. Computational Statistics, 23, 317-335.

[2] Castle, J. L., Doornik, J. A., Hendry, D. F. y Pretis, F. (2015) Detecting location shifts during model selection using step-indication saturation. Econometrics, 3, 240-260.

[3] Chen, C. y Liu, l-M. (1993) Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88, 284-297.

[4] Galeano, P., Peña, D. y Tsay, R. S. (2006) Outlier detection in multivariate time series by projection pursuit. Journal of the American Statistical Association, 101, 654-669.

[5] Galeano, P., Peña, D. y Tsay, R. S. (2022a) Detecting outliers in time series with an orthogonal greedy algorithm. Manuscript.

[6] Galeano, P., Peña, D. y Tsay, R. S. (2022b) Outlier detection in large dimensional time series. Manuscript.

[7] Tsay, R. S., Peña, D. y Pankratz, A. E. (2000) Outliers un multivariate time series. Biometrika, 87, 789-804.

# Métodos robustos para datos funcionales basados en sieves y/o penalizaciones

**Graciela Boente Boente (Universidad de Buenos Aires)**

Como es bien sabido, con los desarrollos tecnológicos, es cada vez más frecuente obtener y almacenar datos registrados continuamente durante un intervalo de tiempo o intermitentemente en varios puntos de tiempo discretos. Ambos son ejemplos de datos funcionales, que se han convertido en un tipo de datos común en muchas áreas como quimiometría, economía, medio ambiente, reconocimiento de imágenes y espectroscopia, entre otros. El análisis de datos funcionales (FDA) abarca la metodología estadística para datos que son intrínsecamente infinito-dimensionales.

En particular, los modelos de regresión funcional postulan una relación entre una respuesta escalar y una covariable funcional, siendo los modelos de regresión lineal semifuncional una extensión de dichos modelos al caso en que se observa además una variable univariada que entra en el modelo en forma noparamétrica. A diferencia del modelo de

regresión lineal cuando las covariables son de dimensión p, la estimación de la pendiente funcional implica resolver un problema que resulta indeterminado si consideramos la extensión directa del caso finito-dimensional. Por ello, para resolverlo debemos recurrir a otras metodologías estadísticas.

A este problema causado por la dimensión, se agrega la sensibilidad a datos atípicos de los estimadores obtenidos por los métodos clásicamente usados ya que se basan en la minimización de la suma de los cuadrados de los residuos. Por esa razón, es de importancia práctica obtener estimadores que sean robustos frente observaciones atípicas tanto verticales como de alta palanca, que generalmente son las más difíciles de identificar y pueden causar graves daños al estimador de mínimos cuadrados. El mismo problema que afecta la estimación del parámetro de regresión aparece cuando intentamos definir estimadores de las direcciones canónicas que es un problema de interés cuando se observan en un mismo individuo dos variables funcionales y se desea proveer una reducción de dimensión con máxima asociación.

En esta charla describiremos algunos procedimientos robustos para analizar observaciones correspondientes a datos funcionales que ilustraremos en ejemplos reales.

Consideraremos asimismo el problema de estimar las direcciones principales cuando hay datos esparsos pues es de interés en el caso de regresión funcional. Los resultados que presentaremos se basan en trabajos realizados y/o en curso con Nadia Kudraszow, Matías Salibián Barrera y Pablo Vena.

## Seroprevalencia de la infección por SARS-CoV-2 en España: Estudio ENE-COVID

**Marina Pollán Santamaría (Instituto de Salud Carlos III)**

### 1. INTRODUCCIÓN

El objetivo del Estudio Nacional sero-Epidemiológico de la infección por SARS-CoV-2 ENE-COVID fue caracterizar la difusión de la epidemia en nuestro país.

### 2. DESARROLLO Y RESULTADOS DEL ESTUDIO

ENE-COVID es un estudio longitudinal de base poblacional en el que los participantes fueron seleccionados mediante muestreo bietápico estratificado por provincia y tamaño municipal, seleccionando de forma aleatoria 1500 secciones censales (1ª etapa) y 24 hogares (2ª etapa) en ellas. Todas las personas presentes en el hogar fueron invitadas a participar en las 4 rondas del estudio [1]. Se utilizaron dos tests de anticuerpos IgG complementarios, validados previamente. Un test rápido (digitopunción), que feacilita

la participación y un inmunoensayo quimioluminiscente de micropartículas que requiere muestra de sangre y presenta mayor precisión [1].

Todos los análisis se realizan asignando a cada participante del estudio un peso de muestreo inversamente proporcional a su probabilidad de selección, ajustado también por la tasa de no respuesta específica según sexo, grupo de edad y nivel de renta relativo de la sección censal [1].

Las 3 primeras rondas del estudio, llevadas a cabo cada 3 semanas desde finales de abril a finales de junio de 2020, informan de la primera onda epidémica [1]. La seroprevalencia se situó en torno al 5 %, con una gran heterogeneidad geográfica: mientras Ceuta, Murcia, Asturias, Galicia, Baleares y Canarias tenían prevalencias inferiores o cercanas al 2 %, las Comunidades de Castilla-La Mancha y Madrid se aproximaban o superaban el 10 % [1, 2]. ENE-COVID proporcionó información de prevalencia de infección en todos los grupos de edad, desde bebés hasta nonagenarios, sin detectar grandes diferencias, salvo una menor prevalencia en niños y adolescentes [2], mientras que la imagen obtenida a través de los casos confirmados en ese momento parecía indicar mayores tasas de infección en los grupos de mayor edad [3]. ENE-COVID también mostró que SARS-CoV-2 infectaba por igual a hombres y mujeres [4], conclusión diferente a la obtenida con la información de casos confirmados. Nuestro estudio cuantificó también la proporción de infecciones asintomáticas, en torno al 30 % [2]. Los datos de seroprevalencia identificaron también el mayor riesgo de infección del personal sanitario [2] y de las personas que convivieron con pacientes COVID-19 o con personas con síntomas compatibles con esta enfermedad [2].

Combinando la información de ENE-COVID con los datos de mortalidad obtenidos en la Red Nacional de Vigilancia Epidemiológica y en el Sistema de Monitorización diaria de la Mortalidad (MoMo) hemos podido cuantificar la letalidad del SARS-CoV-2 en la población española no institucionalizada durante esta primera onda epidémica, situándose entre un 0,8 % y un 1,1 %, con grandes diferencias en función de la edad y el sexo [5].La mortalidad entre los infectados menores de 50 años fue escasa, mientras que fue superior al 10 % en los hombres de 80 años y más y en torno al 5-6 % en las mujeres del mismo grupo etario [5]. Estas cifras, como decimos, no reflejan lo ocurrido en las residencias de mayores.

La 4ª ronda (2ª quincena de noviembre) de ENE-COVID ha servido para caracterizar la segunda onda epidémica [1]. En esas fechas, uno de cada 10 españoles había sido infectado, mientras que en Madrid, Albacete, Cuenca y Soria la seroprevalencia global se aproximaba ya al 20 %. Las diferencias geográficas, todavía muy patentes, se reducían como consecuencia de la penetración del virus en territorios apenas afectados durante la primera onda epidémica. Mientras que en dicha onda apenas se percibían diferencias

según el nivel de renta, en la segunda onda epidémica esas diferencias comienzan a ser visibles [1].

## 3. DISCUSIÓN Y CONCLUSIONES

ENE-COVID ha servido para describir las dos primeras ondas epidémicas en nuestro país, proporcionando información de las dimensiones reales de la pandemia y las características sociodemográficas relacionadas con una mayor o menor probabilidad de infección.

## REFERENCIAS

[1] Instituto de Salud Carlos III. Estudio Nacional de sero-Epidemiología de la infección por SARS-CoV-2 en España (ENECOVID) [Internet]. COVID-19 en España. Disponible en: https://portalcne.isciii.es/enecovid19/.

[2] Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M.A., Pérez-Olmeda, M., et al. (2020) Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet Lond Engl., 396 (10250), 535-44.

[3] CNE-RENAVE (2020) Análisis de los casos de COVID-19 notificados a la RENAVE hasta el 10 de mayo de 2020 en España [Internet]. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Report No.: 33. Disponible en: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/Informe%20n%c2%ba%2033.%20An%c3%a1lisis%20de%20los%20casos%20de%20COVID-19%20hasta%20el%2010%20de%20mayo%20en%20Espa%c3%b1a%20a%2029%20de%20mayo%20de%202020.pdf.

[4] Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Pérez-Olmeda, M., Yotti, R. (2020) SARS-CoV-2 seroprevalence in Spain-Authors' reply. The Lancet, 396(10261), 1484-5.

[5] Pastor-Barriuso, R., Pérez-Gómez, B., Hernán, M.A., Pérez-Olmeda, M., Yotti, R., Oteo-Iglesias, J., et al. (2020) Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: nationwide seroepidemiological study. BMJ, m4509.

# Multivariate conditional transformation models

**Thomas Kneib (University of Göttingen)**

Regression models describing the joint distribution of multivariate response variables conditional on covariate information have become an important aspect of contemporary

regression analysis. However, a limitation of such models is that they often rely on rather simplistic assumptions, e.g. a constant dependence structure that is not allowed to vary with the covariates or the restriction to linear dependence between the responses only. We propose a general framework for multivariate conditional transformation models that overcomes such limitations and describes the full joint distribution in a tractable and interpretable yet flexible way. Among the particular merits of the framework are that it can be embedded into likelihood-based inference (including results on asymptotic normality) and allows the dependence structure to vary with the covariates. In addition, the framework scales well beyond bivariate response situations.

## Índices de poder en juegos simples con estructura de partición

**José María Alonso Meijide (Universidade de Santiago de Compostela)**

La Teoría de Juegos es una rama de las Matemáticas que se dedica al estudio de los problemas de decisión en los que interaccionan varios agentes. La Teoría de Juegos es una herramienta fundamental para las Ciencias Sociales (especialmente para la Economía y la Politología), pero también se ha aplicado con éxito en otros ámbitos, como la Biología, las Ciencias Medioambientales, o los problemas militares.

Los juegos suelen clasificarse en dos tipos: los no cooperativos y los cooperativos. En los juegos cooperativos, un conjunto de jugadores $N$ disponen de mecanismos que les permiten tomar acuerdos vinculantes. El problema central de los juegos cooperativos es el estudio de soluciones, es decir, cómo deben repartirse entre los jugadores los beneficios que se generan con su cooperación. Vamos a distinguir dos tipos de juegos cooperativos: juegos en forma de función característica y juegos en forma de función de partición. En un juego en forma de función característica, dada una coalición (un conjunto de jugadores) la función característica del juego asigna a esta coalición el pago que puede asegurarse, independientemente de cómo actúen el resto de jugadores. El valor de Shapley (1953) es una de las soluciones más estudiadas para los juegos en forma de función característica. Se define como un promedio de lo que aporta un jugador a las coaliciones a las que puede unirse. Un concepto básico es el de jugador nulo, que es aquel que no aporta nada a ninguna coalición, es decir, es un jugador para el que todas las contribuciones son nulas.

Thrall y Lucas (1963) introducen los juegos en forma de partición. En este caso, el pago que recibe una coalición depende de la forma en la que estén organizados el resto de jugadores. Si la función característica está definida sobre el conjunto de todos los subconjuntos de $N$, la función de partición está definida sobre sobre el conjunto de coaliciones integradas (o embebidas), es decir, pares formados por una coalición y la partición

que constituyen los restantes jugadores. Pueden definirse diferentes generalizaciones del valor de Shapley para los juegos en forma de partición dependiendo del concepto de contribución considerado. En este caso, cuando un jugador deja una coalición, tendría la opción de permanecer solo o podría unirse a otro elemento de la partición. Estas diferencias también determinan distintas generalizaciones del concepto de jugador nulo para juegos en forma de partición.

En esta charla, después de introducir algunos elementos básicos de los juegos en forma de partición se revisarán algunos de los resultados que hemos obtenido para juegos en forma de partición (Alonso-Meijide y otros, 2017 y 2019). En particular, se presentarán diversas soluciones que surgen en el caso de los juegos simples en forma de función de partición. En este tipo de juegos, la función en forma de partición solamente toma los valores 0 y 1, dando lugar a coaliciones embebidas perdedoras o ganadoras, respectivamente. En este contexto definimos coaliciones minimales ganadoras y proponemos y caracterizamos dos soluciones (índices de poder) basados en estas coaliciones. Hemos aplicado estos índices para estudiar la distribución del poder en el Parlamento de Andalucía surgido tras las elecciones del 22 de marzo de 2015. Los resultados presentados en esta charla son fruto de la colaboración con Mikel Álvarez Mozos (Universitat de Barcelona), Gloria Fiestras Janeiro (Universidade de Vigo) y Andrés Jiménez Losada (Universidad de Sevilla).

## REFERENCIAS

[1] Alonso-Meijide, J. M, Álvarez-Mozos, M. y Fiestras-Janeiro, M. G. (2017) Power indices and minimal winning coalitions for simple games in partition function form. Group Decision and Negotiation, 26, 1231-1245

[2] Alonso-Meijide, J. M, Álvarez-Mozos, M., Fiestras-Janeiro, M. G. y Jiménez-Losada, A. (2019) Complete null agent for games with externalities. Expert Systems with Applications, 135, 1-11

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

[4] Thrall, R. M., and Lucas, W. F. (1963) $N$-person games in partition function form. Naval Research Logistics Quarterly, 10, 281–298.

# Sesión de Biometría

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Bayesian nonparametric inference for the overlap coefficient: with an application to disease diagnosis

Vanda Inácio[1], Javier E. Garrido Guillén[1]

[1] School of Mathematics, University of Edinburgh, Scotland, UK

## ABSTRACT

Diagnostic tests play an important role in medical research and clinical practice. The ultimate goal of a diagnostic test is to distinguish between diseased and nondiseased individuals and before a test is routinely used in practice, it is a pivotal requirement that its ability to discriminate between these two states is thoroughly assessed. The overlap coefficient, which is defined as the proportion of overlap area between two probability density functions, has gained popularity as a summary measure of diagnostic accuracy. We propose a flexible Bayesian modelling framework, based on Dirichlet process mixtures and the Bayesian bootstrap, for conducting inference about the overlap coefficient. The results of the simulation study show that our estimator is able to recover the true value of the overlap coefficient, and that the empirical coverage probability of the corresponding 95% credible intervals is also close to the nominal value, under a variety of conceivable test outcomes distributions. Our approach is illustrated through an application concerned with the search for biomarkers of ovarian cancer.

**Keywords**: Bayesian bootstrap, diagnostic test, Dirichlet process mixtures, ovarian cancer, overlap coefficient.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

**Estimating sperm whale acoustic cue rates from tags without acoustic data**

Tiago A. Marques[1], Carolina S. Marques[2] and Kalliopi C. Gkikopoulou[3]

[1]Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland & Centro de Estatística e Aplicações, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Portugal
[2]Centro de Estatística e Aplicações, Faculdade de Ciências da Universidade de Lisboa, Portugal
[3]Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland

## ABSTRACT

Density estimation of cetaceans was traditionally conducted using visual methods. In recent years approaches based on using sounds produced by the animals to estimate abundance has gained traction. A possible way of using passive acoustics for density estimation involves estimating a density of sounds and convert it into to a density of animals. To do so we require a cue production rate. Cue production rates are still lacking for many species. The standard way to collect information about sound production rates for cetaceans is to deploy animal-borne tags with acoustic sensors. We present a dataset of acoustic tags placed on sperm whales at a set of 8 locations in 14 different years (in the 2001-2019 period) to investigate how cue production rates change over time and space, and in particular how sound production depends on depth. We investigate two different approaches to estimate cue production rates: (1) a "traditional" approach as was presented by Warren et al. (2017) for beaked whales, using conventional regression models, and (2) a point process framework that might be more suitable since it avoids the need to discretize time into arbitrary time units for modelling. We then present an example of using the derived models to estimate acoustic cue production time-depth recorder tags. Hence, exploiting the relationship between cue production and depth, we estimate acoustic production rates from tags without acoustic data.

This research is part of the ACCURATE project, funded by the US Navy Living Marine Resources program. TAM and CSM thank partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020). We thank the many researchers that made the tag data available.

**Keywords**: abundance, acoustic tags, cue rate, distance sampling, sperm whale.

## REFERENCES

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. (2017) Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation. The Journal of the Acoustical Society of America, 141, 1962–1974.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Rexións de referencia condicionais baseadas en modelos de localización e escala: Aplicación a marcadores glicémicos

Javier Roca Pardiñas[1]

[1] Departamento Estatística e Investigación Operativa, Universidade de Vigo

Moitas decisións clínicas tómanse en base aos resultados dunha única proba diagnóstica. Con todo, para certas enfermidades é necesario o uso de dúas probas diagnósticas. Isto obriga a dispor dunha rexión de referencia bivariada para unha correcta interpretación dos resultados. Ademais, ditas rexións deberán ser axustadas por certas variables (p. ex., idade e sexo). Na actualidade non existen moitos métodos que permitan abordar devandito problema, e nas propostas existentes asúmese distribución gaussiana, que pode ser bastante restritiva en situacións reais. Aquí veremos un método estatístico que permite obter rexións de referencia flexibles axustadas por covariables. En particular, proponse a utilización de nun modelo de localización e escala que permite obter rexións onde se atopa a resposta bivariante cunha determinada probabilidade prefixada de antemán. Finalmente, mostrarase a utilidade desta metodoloxía para obter rexións de confianza (axustadas pola idade do paciente) para dous marcadores usados habitualmente no diagnóstico da diabetes: a hemoglobina glicada e a glicosa plasmática.

# Comunicacións orais

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# EL IMPACTO DEL RENDIMIENTO DE LOS SISTEMAS DE MEDIDA EN LA EVALUACIÓN DE LA CAPACIDAD DEL PROCESO.

Andrés Carrión[1], Angela Grisales[2] y Javier Neira[3]

[1] Centro de Gestión de la Calidad y del Cambio. Universitat Politécnica de Valencia
[2] Centro de Gestión de la Calidad y del Cambio. Universitat Politécnica de Valencia
[3] Depto. de Estadística e IO Aplicadas y Calidad. Universitat Politécnica de Valencia

**RESUMEN**

## Resúmen

La evaluación de la calidad de productos y procesos depende de la posibilidad de obtener mediciones preciosas de un modo consistente en el tiempo. Con esas mediciones se emplearán diferentes herramientas de para monitorizar el rendimiento del proceso, su estabilidad y la conformidad del producto. Una des esas herramientas son los estudios de capacidad, ampliamente conocida y cada vez más considerada en los procesos de toma de decisiones en la industria. Mediante una serie de indicadores, como $C_p$ y $C_{pk}$, se obtiene una información esencial sobre la habilidad de nuestros procesos para cumplir con las especificaciones requeridas. Dado que estos indicadores se calculan a partir de las mediciones obtenidas en procesos de medición, la consistencia de los índices de capacidad depende en gran manera de la calidad y confiabilidad de las mediciones obtenidas. En esta comunicación se analiza la relación entre el rendimiento de los procesos de medida y la representatividad de los estudios de capacidad.

**Palabras y frases clave:** Measurement System Analysis; Capacidad; Control de Calidad; Repetibilidad; Reproducibilidad.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# GRÁFICOS DE CONTROL NO PARAMÉTRICOS PARA LA DISPERSIÓN. OPTIMIZACIÓN CONSIDERANDO ERRORES DE REDONDEO

Javier Porcel-Marí[1], Vicent Giner-Bosch[2], Andrés Carrión-García[2],
Carlos J. Pérez-González[3], Philippe Castagliola[4]

[1] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, València, España
[2] Centro de Gestión de la Calidad y el Cambio, Universitat Politècnica de València, València, España
[3] Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Tenerife, España
[4] Université de Nantes & LS2N UMR CNRS 6004, Nantes, Francia

## RESUMEN

El presente trabajo aborda el diseño óptimo de gráficos de control no paramétricos para la dispersión basados en el estadístico del signo. Se desarrollan e implementan algoritmos de optimización específicos bajo dos supuestos: asumiendo la ausencia, por un lado, y la presencia, por otro, de errores de redondeo debidos a la resolución del instrumento de medida. En concreto, en primer lugar, se realiza una revisión de la literatura, se estudia la distribución en el muestreo del estadístico del signo para la dispersión sin y con errores de redondeo, y se plantea la elaboración de un gráfico de control tipo Shewhart basado en la monitorización de este estadístico. A partir de ahí, se obtienen expresiones analíticas para el cálculo de la tasa de falsas alarmas y la potencia para detectar un cambio en la dispersión del proceso objeto de interés. Se diseña un procedimiento para la determinación óptima de los parámetros del gráfico a partir de los requisitos establecidos por el usuario final; este procedimiento está basado en una estrategia enumerativa implícita. Se llevan a cabo experiencias computacionales con el fin de evaluar la degradación en el comportamiento o desempeño del gráfico diseñado sin tener en cuenta los errores de redondeo en escenarios con presencia de dichos errores, y hasta qué punto la reoptimización teniendo en cuenta los errores de redondeo permite recuperar las propiedades de este. Las pruebas realizadas ponen de manifiesto que, en la mayoría de los casos, la tasa de falsas alarmas y la potencia del gráfico para detectar situaciones de falta de control resultan ser peores de lo esperado en presencia de errores de redondeo, si el gráfico no se diseñó teniéndolos en cuenta. Por el contrario, si el gráfico se diseña desde el inicio considerando la resolución (no infinitesimal) del instrumento de medida, se consiguen valores de eficiencia muy similares a los que teóricamente se podrían haber obtenido en ausencia de errores de redondeo, todo ello respetando el nivel de falsas alarmas estipulado. Las experiencias llevadas a cabo indican, asimismo, la importancia de tener información sobre la forma de la distribución subyacente (en especial, su curtosis) a la hora de optimizar correctamente el gráfico.

**Palabras y frases clave:** Control estadístico de procesos; métodos no paramétricos; estadístico del signo; medidas de dispersión; error de redondeo; distribuciones de Johnson; optimización en ingeniería.

## AGRADECIMIENTOS

Este trabajo ha contado con el apoyo del Ministerio de Ciencia e Innovación de España a través del proyecto PID2019-110442GB-I00.

**REFERENCIAS**

Amin RW, Reynolds MR, Saad B (1995). Nonparametric quality control charts based on the sign statistic. Communications in Statistics - Theory and Methods, 24(6):1597–1623.

Castagliola P, Tran KP, Celano G, Maravelakis PE (2020). The Shewhart Sign Chart with Ties: Performance and Alternatives. En: Koutras M, Triantafyllou I (eds.), Distribution-Free Methods for Statistical Process Monitoring and Control. Springer, Cham.

Chakraborti S, Van der Laan P, Bakir ST (2001). Nonparametric control charts: An overview and some results. Journal of Quality Technology, 33(3):304–315.

Gibbons JD, Chakraborti S (2003). Nonparametric Statistical Inference, 4.ª ed. Dekker, New York.

Johnson NL (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 36(1/2):149–176.

Linna KW, Woodall WH (2001). Effect of measurement error on Shewhart control charts. Journal of Quality Technology, 33(2):213–222.

Lowry CA, Champ CW, Woodall WH (1995). The performance of control charts for monitoring process variation. Communications in Statistics - Simulation and Computation, 24:409-437.

Maleki MR, Amiri A, Castagliola P (2017). Measurement errors in statistical process monitoring: A literature review. Computers & Industrial Engineering, 103:316–329.

Pawar VY, Shirke, DT, Khilare SK (2018). A Nonparametric Control Chart for Process Variability Based on Quantiles. International Journal of Statistics and Economics, 19(3):55–64.

Porcel Marí J (2021). Diseño óptimo de gráficos de control no paramétricos para la dispersión basados en el estadístico de signo en presencia de errores de redondeo. Tesis de Máster, Universitat Politècnica de València.

Riaz M (2008). A Dispersion Control Chart. Communications in Statistics - Simulation and Computation, 37(6):1239–1261.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

## Regresión de Mínimos Cuadrados Parciales (PLS) para datos de respuesta binaria y su representación biplot asociada

Laura Vicente-Gonzalez [1] y Jose Luis Vicente-Villardon[1]

[1]Departamento de Estadística, Universidad de Salamanca

### RESUMEN

En este trabajo se desarrolla una generalización de la Regresión de Mínimos Cuadrados Parciales (PLSR) para analizar datos cuya matriz de respuesta contiene variables binarias y su matriz de predictores es un conjunto de variables numéricas. Este método ha sido denominado Mínimos Cuadrados Parciales para Respuestas Binarias (PLS-BR).

Además del modelo y un algoritmo de ajuste se incluirá una representación biplot asociada y una aplicación a datos reales. La representación biplot asociada es una generalización de la propuesta para respuestas continuas de Oyedele and Lubbe (2015), haciendo una combinación entre la representación biplot tradicional para datos numéricos y un biplot logístico tal y como fue descrito por Vicente-Villardon et al. (2006) o Demey et al. (2008).

Finalmente, para ilustrar la utilidad de este tipo de técnicas se empleará un conjunto de datos reales sobre vinos españoles extraído de Rivas-Gonzalo et al. (1993). El análisis será realizado con el software estadístico R, concretamente con el paquete MultBiplotR (Vicente-Villardon, 2021).

**Palabras y frases clave**: regresión, mínimos cuadrados parciales, datos binarios, biplot

### REFERENCIAS

Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics, 17(3), 166-173.

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., and Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. Bioinformatics, 24(24), 2832-2838.8.

Oyedele, O. F., and Lubbe, S. (2015). The construction of a partial least-squares biplot. Journal of Applied Statistics, 42(11), 2449-2460.

Rivas-Gonzalo, J. C., Gutiérrez, Y., Polanco, A. M., Hebrero, E., Vicente, J. L., Galindo, P., and Santos-Buelga, C. (1993). Biplot analysis applied to enological parameters in the geographical classification of young red wines. American Journal of Enology and Viticulture, 44(3), 302-308.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Vicente-Villardón, J. L., Galindo-Villardón, M. P., and Blázquez-Zaballos, A. (2006). Logistic biplots. Multiple correspondence analysis and related methods. London: Chapman & Hall, 503-521.

Vicente-Villardon, J. L. (2021). MultBiplotR: Multivariate Analysis Using Biplots in R. R package version 1.3.30.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# CONTRIBUTIONS TO PROCESS CONTROL WITH CENSORED DATA ON THE LEFT: APPLICATION OF THE CEV ALGORITHM ON A REAL CASE

Javier Orlando Neira Rueda[1], Andrés Carrión García[2] and Angela Grisales del Río[3]

[1] Depto. de Estadística e IO Aplicadas y Calidad. Universitat Politécnica de Valencia
[2] Centre for Quality and Change Management. Universitat Politecnica de Valencia
[3] Centre for Quality and Change Management. Universitat Politecnica de Valencia

## ABSTRACT

Measuring industrial processes and other practical magnitudes presents problems related with accuracy, variability, and sensitivity. Industrial processes are monitored to detect changes in process or product parameters in order to promptly correct problems that may arise, generating a particular area of scientific interest. This is particularly critical and complex when the measured value falls below the sensitivity limits of the measuring devices available below practical detection limits, causing that part of the observations remain incomplete or unknown. Such observations are referred as incomplete observations or left censored data. With a high level of censorship, for example greater than 60%, the application of traditional methods for monitoring processes is not appropriate. It is required to use appropriate statistical data analysis techniques, to assess the actual state of the process at any time, despite censoring. This document proposes a way to estimate the process parameters in such cases and presents the corresponding control chart, based on an algorithm known as Conditional Expected Value (CEV), also proposing some modifications to improve its preformance.

**Keywords:** Censored data; Control Charts; Process Control; Quality Control.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

## Modelling periodic data with a two-pieces distribution

Jose Ameijeiras-Alonso[1], Irène Gijbels[2] and Anneleen Verhasselt[3]

[1]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Spain
[2]Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Belgium
[3]Center for Statistics, Hasselt University, Belgium

**ABSTRACT**

In this talk, we will illustrate how to model periodic data using circular distributions. In particular, we will show a new family of distributions that is flexible and unimodal. Their associated density functions will be two-pieces in the sense that are defined differently at the left and at the right of the modal direction. We will see how this family of distributions extends many well-known circular distributions, such as the Batschelet and Papakonstantinou models. These densities contain four parameters: modal direction, concentration, peakedness (curvature) at the left and at the right of the modal direction. We will show some properties of the densities belonging to this family of distributions and we will apply them to model a real dataset.

**Keywords**: Circular Statistics, Flexible Modeling, Peakedness, Skewness, Unimodality.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# A TRACTABLE FAMILY OF TESTS OF UNIFORMITY ON THE HYPERSPHERE

Eduardo García-Portugués[1]

[1]Departamento de Estadística, Universidad Carlos III de Madrid

## ABSTRACT

We provide a general and tractable family of tests of uniformity on the hypersphere $\Omega_q = \{\mathbf{x} \in \mathbb{R}^{q+1} : \|\mathbf{x}\| = 1\}$, $q \geq 1$. The family is constructed from powers of chordal distances between pairs of observations. It connects and extends (to arbitrary $q \geq 1$ and with enhanced tractability) three particular tests: Rayleigh (1919), Pycke (2007, 2010), and Bakshaev (2010). The asymptotic null distributions of the new tests are obtained and, despite involving infinite sums of weighted chi-squared random variables, are shown to be tractable. Additionally, powers of the tests against sequences of generic $\sqrt{n}$-alternatives are provided. In particular, explicit powers against novel Cauchy-like distributions on $\Omega_q$, that are of independent interest, are derived. Numerical experiments corroborate the obtained theoretical results. Two real data applications of astronomical and biological nature illustrate the practical use of the tests for assessing uniformity on $\Omega_2$.

**Keywords**: Directional data; Hypersphere; Tests; Uniformity.

## REFERENCES

Bakshaev, A. (2010). N-distance tests of uniformity on the hypersphere. Nonlinear Analysis: Modelling and Control, 15, 15–8.

Pycke, J.-R. (2007). A decomposition for invariant tests of uniformity on the sphere. Proceedings of the American Mathematical Society, 135, 2983–2993.

Pycke, J.-R. (2010). Some tests for uniformity of circular distributions powerful against multimodal alternatives. The Canadian Journal of Statistics, 38, 80–96.

Rayleigh, Lord. (1919). On the problem of random vibrations, and of random flights in one, two, or three dimensions. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 37, 321–347.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# A NEW APPROACH TO DIRECTIONAL CLUSTERING FROM SET ESTIMATION TECHNIQUES

Paula Saavedra-Nieves[1] and Rosa M. Crujeiras[1]

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

## ABSTRACT

Set estimation is focused on the reconstruction of a set (or the estimation of any of its features such as its volume or its boundary) from a random sample of points. Target sets to be estimated may appear in different contexts, but from a distribution-based perspective, level set estimation is a problem of interest. Actually, this theory is also linked to clustering methods: Hartigan (1975) defines the number of population clusters as the number of connected components of density level sets. This topic has received some attention in the literature specially for densities supported on an Euclidean space. However, just as density level sets, this clustering approach can be easily extended to more general settings such as the circle or the sphere.

The rationale for establishing the definition of cluster provided by Hartigan (1975) is quite related with the notion of mode. In fact, several cluster algorithms are based on the detection of modes noting that the number of modes (local maxima of the density function) is rarely smaller than the number of clusters. Nevertheless, the concept of cluster is easier to handle, since it has a global and geometrical nature, whereas the local maxima depend on analytical properties.

In this work, we derive some methodology for estimating the number of directional clusters as the number of connected components of directional level sets. From an empirical perspective, directional level sets are estimated using a nonparametric plug-in reconstruction (see, for instance, Saavedra-Nieves and Crujeiras, 2020). An extensive simulation study shows the performance of this estimator for densities supported on the unit circle and the sphere. Additionally, this methodology is applied to analyse some real data sets.

**Keywords**: Connected components, density level sets, directional data.

## REFERENCES

Hartigan, J. A. (1975) Clustering algorithms. John Wiley & Sons, Inc.

Saavedra-Nieves, P. and Crujeiras, R. M. (2020) Nonparametric estimation of directional highest density regions. arXiv preprint arXiv:2009.08915.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# APPLICATIONS OF KERNEL METHODS TO THE ANALYSIS OF WILDLIFE-VEHICLE COLLISIONS

M. I. Borrajo[1], C. Comas[2] and J. Mateu[3]

[1]Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.
[2]Departamento de Matemáticas. Universitat de Lleida.
[3]Departamento de Matemáticas. Universitat Jaume I de Castellón.

## ABSTRACT

Wildlife-vehicle collisions is a natural disturbance and represents a problem of considerable social and environmental importance. The location of these collisions may be seen as a realisation of an underlying point process defined on the road network.

Point processes are a branch of spatial statistics mainly distinguished by its commonly called "double stochasticity", which refers to the fact that both, the location and the number of events happening, are random. Classically, point processes have been studied in the Euclidean plane or three-dimensional space. However, in the last decade, the increasing detail in spatial data collection, the much more information and accuracy on locations, and real problems such the one we present here, have derived into the need of defining and analysing point processes on networks.

In this work we propose a statistically principled method for kernel smoothing of point process data on a linear network, when the first-order intensity depends on covariates. In particular, we define a consistent kernel estimator for the first-order intensity function, we derive the asymptotic bias and variance of the estimator, and adapt some data-driven bandwidth selectors to estimate the optimal bandwidth. These new procedures are applied to the analysis of wildlife-vehicle collisions through a data set containing the whole road network of Catalonia (North-East of Spain) involving 11790 km of roads and providing the locations of 6590 wildlife-vehicle collision points occurred during the period 2010-2014.

**Keywords**: point processes; linear networks; intensity estimation; covariates; wildlife-vehicle collisions.

## REFERENCES

Borrajo, M. I., Comas, C., Costafreda-Aumedes, S. and Mateu, J. (2021). Stochastic smoothing of point processes for wildlife-vehicle collisions on road networks. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-021-02072-3

Borrajo, M. I., González-Manteiga, W. and Martí?nez-Miranda, M. D. (2020). Bootstrapping kernel intensity estimation for inhomogeneous point processes with spatial covariates. Computational Statistics & Data Analysis, 144. https://doi.org/10.1016/j.csda.2019.106875.

Fuentes-Santos, I., González-Manteiga, W. and Mateu, J. (2016). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. Scandinavian Journal of Statistics, 43(2), 416-435. https://doi.org/10.1111/sjos.12183

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

## Novel effect and goodness-of-fit tests for the concurrent model

Laura Freijeiro González[1], Manuel Febrero Bande[2] and Wenceslao González Manteiga[3]

[1,2,3]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

### ABSTRACT

There are situations where the influence between a variable of interest $Y$ and $p$ covariates $X_1, \ldots, X_p$ is concurrent or point-wise in terms of an argument $t$. Then, the value of $Y(t)$ is only influenced by the covariates value in that instant: $X(t) = (X_1(t), \ldots, X_p(t))$. This gives place to the concurrent model given by the relation $Y(t) = m(t, X(t)) + \varepsilon(t)$, where $m(t, X(t))$ is unknown and $\varepsilon(t)$ is the model error.

However, the estimation of the structure $m(t, X(t))$ is quite tricky in practice, specially when $p$ is large. To alleviate these inconveniences, we propose novel effect and goodness-of-fit tests for the concurrent model. In this way, whereas the effect test is able to perform covariates selection and to reduce the problem dimension, the goodness-of-fit test is developed to check if certain assumption about $m(\cdot)$ structure can be assumed, which can facilitate the model estimation.

To implement both tests, they are rewritten as independence tests and then, innovative dependence measures with good statistical qualities are employed. These are the Hilbert-Schmidt independence criterion of Gretton et al. (2005), the martingale difference divergence of Shao and Zhang (2014) and the conditional distance covariance coefficient of Wang et al. (2015). All of them are modifications of the well-known distance correlation of Székely et al. (2007).

**Keywords**: Concurrent model, covariates selection, effect test, goodness-of-fit test.

### ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministerio de Economía, Industria y Competitividad grant MTM2016-76969-P, Xunta de Galicia Competitive Reference Groups 2017-2020 (ED431C 2017/38) and the Xunta de Galicia grant ED481A-2018/264. Besides, we want to acknowledge to the Supercomputing Center of Galicia (CESGA) for the computational facilities.

### REFERENCES

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E., editors, Algorithmic Learning Theory, pages 63-77, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. Journal of the American Statistical Association, 109(507):1302-131

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6):2769-2794.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. Journal of the American Statistical Association, 110(512):1726-1734. PMID: 26877569.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# DATOS FUNCIONAIS E NEUROIMAXE: APLICABILIDADE E COMPARATIVA CON STATISTICAL PARAMETRIC MAPPING (SPM)

Juan A. Arias-López[1], Carmen Cadarso-Suárez[1] y Pablo Aguiar-Fernández[2,3]

[1] Departamento de Estatística, Análise Matemática, e Optimización, Grupo de Bioestatístca e Ciencia de Datos Biomédicos. Universidade de Santiago de Compostela.

juanantonio.arias.lopez@usc.es

[2] Departamento de Medicina Nuclear e Grupo de Imaxe Molecular, Hospital Clínico Universitario (CHUS) e Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela

[3] Grupo de Imaxe Molecular, Departamento de Psiquiatría, Radioloxía e Saúde Pública do CIMUS, Universidade de Santiago de Compostela

## RESUMO

No eido da neuroimaxe clínica, a detección precisa dos niveles de actividade cerebral dun doente e tamén a súa interpretación son esenciais para un diagnóstico temperán e correcto. Por esta razón, os avances en técnicas estatísticas e computacionais para o análise de datos de neuroimaxe son de gran importancia para o futuro da comprensión das enfermidades e trastornos do cerebro e tamén para o seu diagnóstico e tratamento. En anteriores publicacións discutimos a utilidade das técnicas de datos funcionais para o cálculo da función media dunha imaxe e dos seus intervalos de confianza simultáneos (SCC) tanto para un grupo coma para o cálculo das diferenzas entre dous grupos[1]. En dita publicación centrámonos nos custos computacionais asociados a esta novidosa metodoloxía. No presente traballo, non obstante, utilizaremos imaxes de Tomografía de Emisión por Positróns (PET) con morte neuronal simulada e coñecida (imitando á da enfermidade de Alzheimer) que serán analizados por dúas vías: unha, a tradicional metodoloxía de Statistical Parametric Mapping (SPM); e a nosa proposta, o uso de datos funcionais e SCC. O obxectivo é detectar as áreas que, segundo cada un dos métodos, presentan hipo-actividade neuronal. Posteriormente, comparamos estes resultados cos datos coñecidos da simulación e calculamos a tasa de sensibilidade destas dúas metodoloxías, para 6 rexións de tamaño crecente, e en 5 niveis de hipoactividade simulada distintos. Os resultados indican que o método SCC presenta sensibilidades iguais ou superiores ás de SPM nas configuracións máis sinxelas da análise (i.e. morte neuronal simulada de grande tamaño e alta intensidade), mais amósase claramente superior a SPM nas configuracións da análise máis complexas (i.e. morte neuronal simulada de pequeno tamaño e intensidades reducidas). Estes resultados, esperables mais pendentes de ser demostrados ata o momento, demostran que todavía hai unha grande marxe de melloría nas técnicas de análise dos datos de neuroimaxe de que dispoñemos, abrindo a porta a avances en investigación e aplicabilidade que signifiquen un diagnóstico temperán e tratamento máis eficaz de moitas enfermidades (e.g. Alzheimer, Parkinson...) de grande relevancia e impacto social.

**Palabras e frases chave**: neuroimaxe, datos funcionais, procesamento de imaxes, neurociencia computacional

## REFERENCIAS

1. Arias-López, J. A., Cadarso-Suárez, C., & Aguiar-Fernández, P. (2021, September). Computational Issues in the Application of Functional Data Analysis to Imaging Data. In International Conference on Computational Science and Its Applications (pp. 630-638). Springer, Cham.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Contraste de Especificación para Series de Tiempo Funcionales con Aplicación a Modelos de Difusión

Alejandra López-Pérez[1], Javier Álvarez-Liébana[2], Manuel Febrero-Bande[1] y Wenceslao González-Manteiga[1]

[1]Departamento de Estadística, Análisis Matemático y Optimización. Universidade de Santiago de Compostela.
[2]Departamento de Estadística e Investigación Operativa. Universidad de Oviedo.

## RESUMEN

La metodología de datos funcionales permite la representación de procesos estocásticos en tiempo continuo como secuencias de variables aleatorias en espacios funcionales. El proceso autorregresivo hilbertiano (ARH) juega un papel principal en el modelado de las dinámicas de series de tiempo. En este trabajo se propone un test de bondad de ajuste para la hipótesis nula compuesta de modelos autorregresivos hilbertianos para un determinado orden $z$, ARH($z$). Además, caracterizamos los modelos de difusión como ARH(1) para realizar un test en dos etapas: primero se contrasta si la muestra funcional y sus valores pasados están relacionados mediante un modelo lineal funcional con respuesta funcional y, posteriormente, bajo linealidad, se realiza un F-test funcional. Como ejemplo ilustrativo utilizaremos el modelo de Ornstein–Uhlenbeck para mostrar las propiedades del test en muestras finitas y aplicaremos la metodología en datos reales.

**Palabras y frases clave**: series de tiempo funcionales; test de bondad de ajuste; contrastes de especificación; modelos de difusión; espacios de Hilbert.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

**An effective tool for clustering multivariate time series with an application to financial markets**

Ángel López-Oriona[1] and José A. Vilar[1]

[1]Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, 15071 A Coruña, Spain.

**ABSTRACT**

Clustering of multivariate time series is a central problem in data mining with applications in many fields. Frequently, the clustering target is to identify groups of series generated by the same multivariate stochastic process. Most of the approaches to address this problem include a prior step of dimensionality reduction which may result in a loss of information or consider dissimilarity measures based on correlations and cross-correlations but ignoring the serial dependence structure. We propose a novel approach to measure dissimilarity between multivariate time series aimed at jointly capturing both cross dependence and serial dependence. Specifically, each series is characterized by a set of matrices of estimated quantile cross-spectral densities, where each matrix corresponds to a pair of quantile levels. Then the distance between every couple of series is evaluated by comparing their estimated quantile cross-spectral densities, and the pairwise dissimilarity matrix is taken as starting point to develop a partitioning around medoids algorithm. Since the quantile-based cross-spectra capture dependence in quantiles of the joint distribution, the proposed metric has a high capability to discriminate between high-level dependence structures. A simulation study shows that our clustering procedure outperforms a wide range of alternative methods and exhibits robustness to noise distribution besides being computationally efficient. A real data application involving bivariate financial time series illustrates the usefulness of the proposed approach.

**Keywords**: Multivariate time series, clustering, dissimilarity measure, quantile cross-spectral density.

## 1 Introduction

Time series clustering is a central problem in data mining with applications in many fields. The objective is to split a large set of unlabeled time series realizations into homogeneous groups so that similar series are placed together in the same group and dissimilar series are located in different groups. This unsupervised classification process is useful to characterize different dynamic patterns without the need to analyze and model each single time series, which is computationally intensive and often far from being the real target. Many methods to cluster time series have been proposed in the literature. Comprehensive overviews including current advances, future prospects, significant references and specific application areas are provided by Liao (2005); Fu (2011); Aghabozorgi, Shirkhorshidi, and Wah (2015), and more recently in the monograph by Maharaj, D'Urso, and Caiado (2019). However, most of the proposed approaches concern univariate time series (UTS) while clustering of multivariate time series (MTS) has received much less attention. Unlike UTS, MTS involve a number of variables which must be jointly considered to characterize the underlying dynamic pattern. From the clustering point of view, this is a challenging issue because a dissimilarity measure between MTS should take into account the interdependence relationship between variables.

The focus of this paper is on clustering of MTS according to the underlying dependence structures, i.e., on identifying groups of MTS generated by the same multivariate stochastic process. To that aim, we introduce a metric addressing jointly both cross dependence and serial dependence, which is based on the so-called quantile cross-spectral density. Our experimental analyses show that the proposed measure produces excellent results in clustering of MTS using the PAM algorithm. Compared to other alternative dissimilarities, our metric is clearly more effective in scenarios involving complex dependence models and highly competitive in conventional setups where the cross-spectral density fully characterizes the underlying dependence. Our approach also exhibits robustness to the presence of heavy-tailed noise distributions and high computational efficiency.

The rest of the paper is organized as follows. The dissimilarity measure between MTS based on estimated quantile cross-spectral densities is presented in Section 2. A precise definition of the quantile

cross-spectral density is provided, the estimation procedure is detailed and the corresponding distance measure is defined. Its behaviour in MTS clustering is analyzed in Section 3 throughout a simulation study where different clustering scenarios featured by the kind of generating processes are considered. Effects of different distributional forms for the errors are examined and the results are compared with the ones obtained using other dissimilarity measures. Finally, in Section 4, we apply the proposed tool to cluster real bivariate time series belonging to the field of Finance.

## 2 A novel structure-based approach for multivariate time series clustering

Consider a set of $s$ multivariate time series $\mathcal{S} = \left\{\boldsymbol{X}_t^{(1)}, \ldots, \boldsymbol{X}_t^{(s)}\right\}$, where the $j$-th element $\boldsymbol{X}_t^{(j)} = \left\{\boldsymbol{X}_1^{(j)}, \ldots, \boldsymbol{X}_{T_j}^{(j)}\right\}$ is a $T_j$-length partial realization from any $d$-variate real-valued strictly stationary stochastic process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$. We wish to perform clustering on the elements of $\mathcal{S}$ in such a way that the series generated from the same stochastic process are grouped together. We propose to use a partitional algorithm starting from a pairwise dissimilarity matrix based on comparing estimated quantile cross-spectral densities. In this section, the quantile cross-spectral density notion is presented and then used to define a distance between MTS.

### 2.1 The quantile cross-spectral density

Let $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\} = \{(X_{t,1}, \ldots, X_{t,d}),\, t \in \mathbb{Z}\}$ be a $d$-variate real-valued strictly stationary stochastic process. Denote by $F_j$ the marginal distribution function of $X_{t,j}$, $j = 1, \ldots, d$, and by $q_j(\tau) = F_j^{-1}(\tau)$, $\tau \in [0,1]$, the corresponding quantile function. Fixed $l \in \mathbb{Z}$ and an arbitrary couple of quantile levels $(\tau, \tau') \in [0,1]^2$, consider the cross-covariance of the indicator functions $I\{X_{t,j_1} \leq q_{j_1}(\tau)\}$ and $I\{X_{t+l,j_2} \leq q_{j_2}(\tau')\}$ given by

$$\gamma_{j_1,j_2}(l,\tau,\tau') = \mathrm{Cov}\left(I\{X_{t,j_1} \leq q_{j_1}(\tau)\}, I\left\{X_{t+l,j_2} \leq q_{j_2}(\tau')\right\}\right), \tag{1}$$

for $1 \leq j_1, j_2 \leq d$. Taking $j_1 = j_2 = j$, the function $\gamma_{j,j}(l,\tau,\tau')$, with $(\tau,\tau') \in [0,1]^2$, so-called QAF of lag $l$, generalizes the traditional autocovariance function. While autocovariances measure linear dependence between different lags evaluating covariability with respect to the average, quantile autocovariances examine how a part of the range of variation of $X_j$ helps to predict whether the series will be below quantiles in a future time. This way, QAF entirely describes the dependence structure of $(X_{t,j}, X_{t+l,j})$, enabling us to capture serial features that standard autocovariances cannot detect. Note that $\gamma_{j_1,j_2}(l,\tau,\tau')$ always exists since no assumptions about moments are required. Furthermore, QAF also takes advantage of the local distributional properties inherent to the quantile methods, including robustness against heavy tails, dependence in the extremes and changes in the conditional shapes (skewness, kurtosis). Motivated by these nice properties, a dissimilarity between UTS based on comparing estimated quantile autocovariances over a common range of quantiles was proposed by Lafuente-Rego and Vilar (2016) to perform UTS clustering with very satisfactory results.

In the case of the multivariate process $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\}$, we can consider the $d \times d$ matrix

$$\boldsymbol{\Gamma}(l,\tau,\tau') = \left(\gamma_{j_1,j_2}(l,\tau,\tau')\right)_{1\leq j_1,j_2\leq d}, \tag{2}$$

which jointly provides information about both the cross-dependence (when $j_1 \neq j_2$) and the serial dependence (because the lag $l$ is considered). To obtain a much richer picture of the underlying dependence structure, $\boldsymbol{\Gamma}(l,\tau,\tau')$ can be computed over a range of prefixed values of $L$ lags, $\mathcal{L} = \{l_1, \ldots, l_L\}$, and $r$ quantile levels, $\mathcal{T} = \{\tau_1, \ldots, \tau_r\}$, thus having available the set of matrices

$$\boldsymbol{\Gamma}_{\boldsymbol{X}_t}(\mathcal{L},\mathcal{T}) = \left\{\boldsymbol{\Gamma}(l,\tau,\tau'),\ l \in \mathcal{L},\ \tau,\tau' \in \mathcal{T}\right\}. \tag{3}$$

In the same way as the spectral density is the representation in the frequency domain of the autocovariance function, the spectral counterpart for the cross-covariances $\gamma_{j_1,j_2}(l,\tau,\tau')$ can be introduced. Under suitable summability conditions (mixing conditions), the Fourier transform of the cross-covariances is well-defined and the *quantile cross-spectral density* is given by

$$\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau') = (1/2\pi)\sum_{l=-\infty}^{\infty}\gamma_{j_1,j_2}(l,\tau,\tau')e^{-il\omega}, \tag{4}$$

for $1 \leq j_1, j_2 \leq d$, $\omega \in \mathbb{R}$ and $\tau,\tau' \in [0,1]$. Note that $\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau')$ is complex-valued so that it can be represented in terms of its real and imaginary parts, which will be denoted by $\Re(\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau'))$ and $\Im(\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau'))$, respectively. The quantity $\Re(\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau'))$ is known as quantile cospectrum of $(X_{t,j_1})_{t\in\mathbb{Z}}$ and $(X_{t,j_2})_{t\in\mathbb{Z}}$, whereas the quantity -$\Im(\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau'))$ is called quantile quadrature spectrum of $(X_{t,j_1})_{t\in\mathbb{Z}}$ and $(X_{t,j_2})_{t\in\mathbb{Z}}$.

For fixed quantile levels $(\tau, \tau')$, the quantile cross-spectral density is the cross-spectral density of the bivariate process

$$(I\{X_{t,j_1} \leq q_{j_1}(\tau)\}, I\{X_{t,j_2} \leq q_{j_2}(\tau')\}). \tag{5}$$

Therefore the quantile cross-spectral density measures dependence between two components of the multivariate process in different ranges of their joint distribution and across frequencies. Proceeding as in (3), the quantile cross-spectral density can be evaluated on a range of frequencies $\Omega$ and of quantile levels $\mathcal{T}$ for every couple of components in order to obtain a complete representation of the process, i.e., consider the set of matrices

$$\mathfrak{f}_{\boldsymbol{X}_t}(\Omega, \mathcal{T}) = \left\{\mathfrak{f}(\omega, \tau, \tau'),\ \omega \in \Omega,\ \tau, \tau' \in \mathcal{T}\right\}, \tag{6}$$

where $\mathfrak{f}(\omega, \tau, \tau')$ denotes the $d \times d$ matrix in $\mathbb{C}$

$$\mathfrak{f}(\omega, \tau, \tau') = \left(\mathfrak{f}_{j_1,j_2}(\omega, \tau, \tau')\right)_{1 \leq j_1, j_2 \leq d}. \tag{7}$$

Representing $\boldsymbol{X}_t$ through $\mathfrak{f}_{\boldsymbol{X}_t}$, a complete information on the general dependence structure of the process is available. Comprehensive discussions about the nice properties of the quantile cross-spectral density are given in Lee and Rao (2012), Dette, Hallin, Kley, and Volgushev (2015) and Baruník and Kley (2019), including invariance to monotone transformations, robustness and capability to detect nonlinear dependence. It is also worth enhancing that the quantile cross-spectral density provides a full description of all copulas of pairs of components in $\boldsymbol{X}_t$, since the difference between the copula of an arbitrary couple $(X_{t,j_1}, X_{t+l,j_2})$ evaluated in $(\tau, \tau')$ and the independence copula at $(\tau, \tau')$ can be written as

$$\mathbb{P}\left(X_{t,j_1} \leq q_{j_1}(\tau), X_{t+l,j_2} \leq q_{j_2}(\tau')\right) - \tau\tau' = \int_{-\pi}^{\pi} \mathfrak{f}_{j_1,j_2}(\omega, \tau, \tau') e^{il\omega}\, d\omega.$$

According with the prior arguments, a dissimilarity measure between realizations of two multivariate processes, $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, could be established by comparing their representations in terms of the quantile cross-spectral density matrices, $\mathfrak{f}_{\boldsymbol{X}_t}$ and $\mathfrak{f}_{\boldsymbol{Y}_t}$, respectively. For it, estimates of the quantile cross-spectral densities must be obtained.

Let $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T\}$ be a realization from the process $(\boldsymbol{X}_t)_{t \in \mathbb{Z}}$ so that $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,d})$, $t = 1, \ldots, T$. For arbitrary $j_1, j_2 \in \{1, \ldots, d\}$ and $(\tau, \tau') \in [0,1]^2$, Baruník and Kley (2019) propose to estimate $\mathfrak{f}_{j_1,j_2}(\omega, \tau, \tau')$ considering a smoother of the cross-periodograms based on the indicator functions $I\{\hat{F}_{T,j}(X_{t,j})\}$, where $\hat{F}_{T,j}(x) = T^{-1}\sum_{t=1}^{T} I\{X_{t,j} \leq x\}$ denotes the empirical distribution function of $X_{t,j}$. This approach extends to the multivariate case the estimator proposed by Kley, Volgushev, Dette, and Hallin (2016) in the univariate setting. More specifically, the called *rank-based copula cross periodogram* (CCR-periodogram) is defined by

$$I_{T,R}^{j_1,j_2}(\omega, \tau, \tau') = \frac{1}{2\pi T} d_{T,R}^{j_1}(\omega, \tau) d_{T,R}^{j_2}(-\omega, \tau'), \tag{8}$$

where

$$d_{T,R}^{j}(\omega, \tau) = \sum_{t=1}^{T} I\{\hat{F}_{T,j}(X_{t,j}) \leq \tau\} e^{-i\omega t}.$$

The asymptotic properties of the CCR-periodogram are established in Proposition S4.1 of Baruník and Kley (2019). Likewise the standard cross-periodogram, the CCR-periodogram is not a consistent estimator of $\mathfrak{f}_{j_1,j_2}(\omega, \tau, \tau')$. To achieve consistency, the CCR-periodogram ordinates (evaluated on the Fourier frequencies) are convolved with weighting functions $W_T(\cdot)$. The *smoothed CCR-periodogram* takes the form

$$\hat{G}_{T,R}^{j_1,j_2}(\omega, \tau, \tau') = (2\pi/T) \sum_{s=1}^{T-1} W_T\left(\omega - \frac{2\pi s}{T}\right) I_{T,R}^{j_1,j_2}\left(\frac{2\pi s}{T}, \tau, \tau'\right), \tag{9}$$

where

$$W_T(u) = \sum_{v=-\infty}^{\infty} (1/h_T) W\left(\frac{u + 2\pi v}{h_T}\right),$$

with $h_T > 0$ a sequence of bandwidths such that $h_T \to 0$ and $Th_T \to \infty$ as $T \to \infty$, and $W$ is a real-valued, even weight function with support $[-\pi, \pi]$. Consistency and asymptotic performance of the smoothed CCR-periodogram $\hat{G}_{T,R}^{j_1,j_2}(\omega, \tau, \tau')$ are established in Theorem S4.1 of Baruník and Kley (2019).

This way, the set of complex-valued matrices $\mathfrak{f}_{\boldsymbol{X}_t}(\Omega, \mathcal{T})$ in (6) characterizing the underlying process can be estimated by

$$\hat{\mathfrak{f}}_{\boldsymbol{X}_t}(\Omega, \mathcal{T}) = \left\{\hat{\mathfrak{f}}(\omega, \tau, \tau'),\ \omega \in \Omega,\ \tau, \tau' \in \mathcal{T}\right\}, \tag{10}$$

where $\hat{\mathbf{f}}(\omega, \tau, \tau')$ is the matrix

$$\hat{\mathbf{f}}(\omega, \tau, \tau') = \left(\hat{G}_{T,R}^{j_1,j_2}(\omega, \tau, \tau')\right)_{1 \leq j_1, j_2 \leq d}. \tag{11}$$

Throughout this article, the smoothed CCR-periodograms were obtained by using the R-package **quantspec** (Kley, 2016).

### 2.2 An innovative spectral dissimilarity measure between MTS

A simple dissimilarity criterion between a pair of $d$-variate time series $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ can be obtained by comparing their estimated sets of complex-valued matrices $\hat{\mathbf{f}}_{\boldsymbol{X}_t^{(1)}}(\Omega, \mathcal{T})$ and $\hat{\mathbf{f}}_{\boldsymbol{X}_t^{(2)}}(\Omega, \mathcal{T})$ evaluated on a common range of frequencies and quantile levels. Specifically, each time series $\boldsymbol{X}_t^{(u)}$, $u = 1, 2$, is characterized by means of a set of $d^2$ vectors $\{\boldsymbol{\Psi}_{j_1,j_2}^{(u)}, 1 \leq j_1, j_2 \leq d\}$ constructed as follows. For a given set of $K$ different frequencies $\Omega = \{\omega_1, \ldots, \omega_K\}$, and $r$ quantile levels $\mathcal{T} = \{\tau_1, \ldots, \tau_r\}$, each vector $\boldsymbol{\Psi}_{j_1,j_2}^{(u)}$ is given by

$$\boldsymbol{\Psi}_{j_1,j_2}^{(u)} = (\boldsymbol{\Psi}_{1,j_1,j_2}^{(u)}, \ldots, \boldsymbol{\Psi}_{K,j_1,j_2}^{(u)}), \tag{12}$$

where each $\boldsymbol{\Psi}_{k,j_1,j_2}^{(u)}$, $k = 1, \ldots, K$ consists of a vector of length $r^2$ formed by rearranging by rows the elements of the matrix

$$(\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'}); i, i' = 1, \ldots, r). \tag{13}$$

Once the set of $d^2$ vectors $\boldsymbol{\Psi}_{j_1,j_2}^{(u)}$ is obtained, they are all concatenated in a vector $\boldsymbol{\Psi}^{(u)}$ in the same way as vectors $\boldsymbol{\Psi}_{k,j_1,j_2}^{(u)}$ constitute $\boldsymbol{\Psi}_{j_1,j_2}^{(u)}$ in (12). In this manner, the dissimilarity between $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ is obtained by means of the Euclidean distance between $\boldsymbol{\Psi}^{(1)}$ and $\boldsymbol{\Psi}^{(2)}$

$$\begin{aligned} d_{QCD}(\boldsymbol{X}_t^{(1)}, \boldsymbol{X}_t^{(2)}) &= \left[||\Re_v(\boldsymbol{\Psi}^{(1)}) - \Re_v(\boldsymbol{\Psi}^{(2)})||^2 + ||\Im_v(\boldsymbol{\Psi}^{(1)}) - \Im_v(\boldsymbol{\Psi}^{(2)})||^2\right]^{1/2} = \\ &\left[\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sum_{i=1}^{r}\sum_{i'=1}^{r}\sum_{k=1}^{K}\left(\Re(\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'})^{(1)}) - \Re(\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'})^{(2)})\right)^2 + \right. \\ &\left.\sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\sum_{i=1}^{r}\sum_{i'=1}^{r}\sum_{k=1}^{K}\left(\Im(\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'})^{(1)}) - \Im(\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'})^{(2)})\right)^2\right]^{1/2}, \end{aligned} \tag{14}$$

where $\Re_v$ and $\Im_v$ denote the element-wise real and imaginary part operations, respectively, and $\hat{G}_{T,R}^{j_1,j_2}(\omega_k, \tau_i, \tau_{i'})^{(u)}$ is the corresponding element of the matrix given by (11) for the series $\boldsymbol{X}_t^{(u)}$.

Below we give some properties illustrating the high ability of $d_{QCD}$ to distinguish between arbitrary differences in the dependence structures between two stochastic processes.

We assume that $\boldsymbol{X}_t^i$ is a $d$-variate, real-valued, strictly stationary process and $\boldsymbol{X}_t^{(i)}$ is a realization of length $T$ from the process $\boldsymbol{X}_t^i$. The $j$-th component of $\boldsymbol{X}_t^i$, $j = 1, \ldots, d$, is denoted by $X_{t,j}^i$. The notation $F_j^i$ stands for the marginal cumulative distribution function of $X_{t,j}^i$. Given a lag $l \in \mathbb{Z}$ and a couple of components $j_1, j_2 = 1, \ldots, d$, the joint cumulative distribution function of the pair $(X_{t,j_1}^i, X_{t+l,j_2}^i)$ is denoted by $F_{j_1,j_2,l}^i$. We suppose that all the mentioned cumulative distribution functions are continuous functions.

**Property 1**. If $\boldsymbol{X}_t^1 = \boldsymbol{X}_t^2$. Then $d_{QCD}(\boldsymbol{X}_t^{(1)}, \boldsymbol{X}_t^{(2)}) \underset{p}{\to} 0$ as $T \to \infty$, where the notation $\underset{p}{\to}$ stands for convergence in probability.

**Property 2**. Assume that there exists some $l \in \mathbb{Z}$ and a couple of dimensions $j_1, j_2 = 1, \ldots, d$ such that $F_{j_1,j_2,l}^1 \neq F_{j_1,j_2,l}^2$ and that $F_j^1 = F_j^2$, $j = 1, \ldots, d$. Then there exist an infinite number of probability levels and an infinite number of frequencies such that $d_{QCD}(\boldsymbol{X}_t^{(1)}, \boldsymbol{X}_t^{(2)}) \underset{p}{\to} a$, $a \neq 0$ as $T \to \infty$.

The proofs of the previous properties can be seen in López-Oriona, Vilar, et al. (2021).

## 3 Experimental evaluation of the proposed clustering procedure

In this section we carry out a set of simulations with the aim of assessing the performance of $d_{QCD}$ in different scenarios of MTS clustering. Firstly we describe the simulation mechanism, then we explain how the assessment of the proposed approach was done and finally we show the results of the simulation study.

### 3.1 Experimental design

The simulated scenarios cover a wide variety of generating processes. Specifically, three unsupervised classification setups were considered, namely clustering of (1) VARMA processes, (2) dynamic conditional correlation processes, and (3) processes exhibiting different types of quantile dependence. The selection of such kind of processes was made with the goal of performing the assessment task in a fair and general manner. Indeed, the three chosen setups are pivotal in several application domains. The generating models concerning each class of processes are given below.

**Scenario 1**. VARMA processes clustering.

(a) VAR(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.5 & 0 \\ -0.4 & 0.5 & 0.3 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \end{pmatrix},$$

(b) VAR(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.4 & 0 \\ -0.4 & 0.5 & 0.4 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \end{pmatrix},$$

(c) VMA(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.5 & 0 \\ -0.4 & 0.5 & 0.3 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} \epsilon_{t-1,1} \\ \epsilon_{t-1,2} \\ \epsilon_{t-1,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \end{pmatrix},$$

(d) VMA(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.4 & 0 \\ -0.4 & 0.5 & 0.4 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} \epsilon_{t-1,1} \\ \epsilon_{t-1,2} \\ \epsilon_{t-1,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \end{pmatrix},$$

(e) VARMA(1,1)

$$\begin{aligned} \begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} &= \begin{pmatrix} 0.6 & 0.5 & 0 \\ -0.4 & 0.5 & 0.3 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \end{pmatrix} + \\ &\quad \begin{pmatrix} 0.6 & 0.5 & 0 \\ -0.4 & 0.5 & 0.3 \\ 0 & -0.5 & 0.7 \end{pmatrix} \begin{pmatrix} \epsilon_{t-1,1} \\ \epsilon_{t-1,2} \\ \epsilon_{t-1,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \end{pmatrix}, \end{aligned}$$

where, in all cases, $(\epsilon_{t,1}, \epsilon_{t,2}, \epsilon_{t,3})^\intercal$ is an i.i.d. vector error process following the trivariate normal distribution with zero mean and covariance matrix equals the identity matrix.

**Scenario 2**. Dynamic conditional correlation processes clustering. Consider $(X_{t,1}, X_{t,2})^\intercal = (a_{t,1}, a_{t,2})^\intercal = (\sigma_{t,1}\epsilon_{t,1}, \sigma_{t,2}\epsilon_{t,2})^\intercal$. The data-generating process consists of two Gaussian GARCH models (Bollerslev, 1986), one which is highly persistent and the other which is not.

$$\begin{aligned} \sigma^2_{t,1} &= 0.01 + 0.05a^2_{t-1,1} + 0.94\sigma^2_{t-1,1}, \\ \sigma^2_{t,2} &= 0.5 + 0.2a^2_{t-1,2} + 0.5\sigma^2_{t-1,2}, \\ \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix} &\sim N\left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix} \right]. \end{aligned}$$

The correlation between the standardized shocks, $\rho_t$, is given by the following expressions:

(a) Constant correlation

$$\rho_t = 0.5,$$

(b) Piecewise constant correlation

$$\rho_t = 0.7I_{\{t \leq (T/2)\}} - 0.9I_{\{t > (T/2)\}},$$

(c) Piecewise constant correlation

$$\rho_t = 0.9 I_{\{t \leq (T/2)\}} - 0.7 I_{\{t > (T/2)\}},$$

(d) Piecewise varying correlation

$$\rho_t = \frac{0.99}{\log(t+2)} I_{\{t \text{ odd}\}} - \frac{0.99}{\log(t+2)} I_{\{t \text{ even}\}},$$

where $I$ stands for the indicator function.

**Scenario 3**. QVAR processes clustering. Consider $= (U_{t,1}, U_{t,2})^\intercal$ a sequence of independent random vectors with independent components $U_{t,k}$ which are uniformly distributed on $[0,1]$, and $\Phi^{-1}(u)$, $u \in (0,1)$, the quantile function of the standard normal distribution.

(a) QVAR(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} 0 & -0.5(U_{t,1} - 0.5) \\ -0.5(U_{t,2} - 0.5) & 0 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} \Phi^{-1}(U_{t,1}) \\ \Phi^{-1}(U_{t,2}) \end{pmatrix},$$

(b) QVAR(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} 0 & 0.5(U_{t,1} - 0.5) \\ 0.5(U_{t,2} - 0.5) & 0 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} \Phi^{-1}(U_{t,1}) \\ \Phi^{-1}(U_{t,2}) \end{pmatrix},$$

(c) QVAR(1)

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} 0 & 1.5(U_{t,1} - 0.5) \\ 1.5(U_{t,2} - 0.5) & 0 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} \Phi^{-1}(U_{t,1}) \\ \Phi^{-1}(U_{t,2}) \end{pmatrix}.$$

The simulation study was carried out as follows. For each scenario, five time series of length $T = 1000$ were generated from each model in order to perform clustering. The distance $d_{QCD}$ between each pair of MTS was calculated using $r = 3$ quantiles of levels 0.1, 0.5 and 0.9 along with the set of Fourier frequencies $\{\omega_k = 2\pi k/T,\ 0 \leq k \leq T/2\}$. The obtained pairwise dissimilarity matrix was then processed by the PAM algorithm to reach the clustering solution. This simulation procedure was performed 100 times for each scenario and each value of $T$.

### 3.2 Alternative metrics and assessment criteria

To shed light on the performance of $d_{QCD}$, clustering solutions based on some state of the art approaches measuring dissimilarity between MTS were also obtained. The considered dissimilarities are summarized below.

- *Dynamic time warping-based distances.* Particularly, the two multivariate extensions of dynamic time warping discussed in Shokoohi-Yekta, Hu, Jin, Wang, and Keogh (2017). The "independent" warping version ($d_{DTWI}$) computes the classical dynamic time warping between each pair of univariate time series, whereas the "dependent" version ($d_{DTWD}$) forces all dimensions to warp identically, in a single warping matrix.

- *Model-based distance.* Specifically, the distance proposed by Maharaj (1999) assessing the difference between vector autoregressive parameter estimates of the series. The original approach implies three main steps. First, a finite order VAR model is fitted to each series via a given criterion. Second, a $p$-value is obtained for each pair of series, regarding the null hypothesis stating that there is no significant difference between both underlying generating processes. Finally, hierarchical clustering is applied to the set of MTS via the $p$-values, but only those series whose associated $p$-value is greater than some predetermined number (e.g 0.05 or 0.01) are grouped together. This implies that the number of clusters is determined by the outcome of the tests of hypotheses and therefore the desired number of groups can not be set in advance. In this case, to make homogeneous comparisons, we simply used the Euclidean distance between the vectors of coefficient estimates of the VAR models ($d_{VAR}$) in the first step to construct the initial dissimilarity matrix. The number of considered lags was determined by the maximum fitted order amongst the MTS as given by the Akaike Information Criterion. This way, when computing the distance between two vectors of coefficients of unequal length, the shortest vector was padded with zeros until it reached the length of the longest vector.

- *PCA-based distance.* Singhal and Seborg (2005) consider a weighted similarity factor based on principal components and the angles between the principal components subspaces. Then, a dissimilarity measure ($d_{PCA}$) is obtained by subtracting the similarity factor from one. In order to perform PCA, we applied the singular value decomposition to the correlation matrices and considered a number of

principal components, $r$, as the minimum value such that at least 95% of the variability of all MTS was explained by means of the first $r$ principal components. This criterion led always to the retention of all principal components.

- *Wavelet-based distance.* D'Urso and Maharaj (2012) introduce a Euclidean distance between wavelet features of MTS, specifically between estimates of wavelet variances and wavelet correlations ($d_W$). The estimates are obtained through the maximum overlap discrete wavelet transform, which requires choosing a wavelet filter of a given length and a number of scales. After performing some brief preliminary analyses, we reached the conclusion that the wavelet filter of length 4 of the Daubechies family, DB4, along with the maximum allowable number of scales, 5 for $T = 250$ and 7 for $T = 1000$, were the choices that led to the best average results in Scenarios 1, 2 and 3. Hence, they were the hyperparameters chosen for the simulation study.

- *Generalized cross correlation-based distance.* Alonso and Peña (2019) propose to measure similarity between two UTS, $X_t$ and $Y_t$, via the generalized cross correlation ($GCC(X_t, Y_t)$), which compares the determinant of the correlation matrix until some lag $r$ of the bivariate vector with those of the two univariate time series. Conceptually, $GCC(X_t, Y_t)$ evaluates the level of linear dependency among both series. Considering the sample correlation matrices, a matrix of pairwise distances of the form $1 - \widehat{GCC}(X_t, Y_t)$ can be directly constructed from every couple of series subjected to the clustering procedure. We propose to extend their approach to a multivariate framework as follows. Each $d$-variate series is first represented through a matrix $\mathbf{X} = (X_{t,j})_{1 \leq t \leq T,\ 1 \leq j \leq d}$, and then described by means of a vector of length $d(d-1)$ whose components are given by

$$1 - \widehat{GCC}(X_{:,j_1}, X_{:,j_2}),\ j_1, j_2 \in \{1, \ldots, d\},\ j_1 \neq j_2, \tag{15}$$

  where $X_{:,i}$ is the $i$-th column of the matrix $\mathbf{X}$. Then we construct a distance matrix by considering the Euclidean distance between the vectors of features ($d_{GCC}$) computed by (15). It is important to remark that the primary goal of Alonso and Peña (2019) is to cluster UTS by linear dependence, i.e. two series are grouped together if they present a high degree of dependence, but the purpose of our extension is different. In our case, the degree of dependence between each pair of UTS within the MTS is evaluated in order to characterize the MTS. In our experiments, the hyperparameter $r$ was set to $r = 1$, a reasonable choice since all models in Scenarios 1, 2 and 3 present one significant lag.

- *Nonparametric dissimilarity in the frequency domain* (Kakizawa, Shumway, & Taniguchi, 1998). A dissimilarity between estimates of the spectral density matrices via the smoothed periodogram. The J-divergence was used to compute the distance between estimates of the spectral density matrices ($d_J$).

The quality of the clustering procedure was assessed by comparing the clustering partition given by the algorithm, $P_k$, with the true cluster solution, which is usually referred to as ground truth, $G_k$. The ground truth consisted of $k = 5$ groups in Scenario 1, $k = 4$ groups in Scenario 2, and $k = 3$ groups in Scenario 3, each one of them involving five time series with the same generating process. The value of $k$ was given as an input parameter to the PAM algorithm. Partitions $P_k$ and $G_k$ were then compared by using three well-known external clustering validity indexes: the Larsen-Aone index (LA) (Larsen & Aone, 1999), the adjusted Rand index (ARI) (Hubert & Arabie, 1985), and the Jaccard index (JI). It is worth remarking that the expected value of ARI is zero for two partitions picked at random according to the generalized hypergeometric distribution. Hence, the value of zero can be associated with a noninformative clustering solution. Additionally, we have also computed a fourth index by considering the 1-nearest neighbour classifier evaluated by leave-one-out cross-validation (LOO1NN). Specifically, LOO1NN index returns the proportion of series which, according to $G_k$, are in the same cluster that their nearest series, based on the given dissimilarity measure. Notice that LOO1NN does not evaluate the clustering algorithm, but gives insights into the quality of the dissimilarity measure. This evaluation criterion has been extensively used in a broad range of pattern recognition applications, including time series clustering (see e.g. Keogh & Kasetty, 2003 and Lafuente-Rego & Vilar, 2016). The indexes LA, JI, and LOO1NN take values between 0 and 1. As for ARI index, it takes values between -1 and 1. In all cases, the closer to one the index, the better the clustering solution.

### 3.3 Results and discussion

Averages and standard deviations of the quality indexes over the 100 trials for the best performing metrics, namely $d_{QCD}$, $d_W$, $d_{GCC}$, $d_J$, $d_{VAR}$ and $d_{PCA}$, are given in Table 1.

According to results in Table 1, the dissimilarity based on the quantile cross-spectral density $d_{QCD}$ produced by far the highest average scores in Scenarios 2 and 3, and presented worse behaviour in Scenario

1. In this scenario, the metrics $d_J$ and $d_{VAR}$ attained the best scores, which was expected because this metrics are mainly aimed at distinguishing between linear processes. When some departures from linearity were taken into account (Scenarios 2 and 3), these dissimilarities substantially decreased their performance, specially $d_{VAR}$, whereas the $QCD$-based measure clearly displayed its capability to differentiate between complex dependence structures. Specifically, in the case of Scenario 3, this metric was the only one capable of totally detecting the true clustering structure in the data. With respect to the metrics $d_W$, $d_{GCC}$ and $d_{PCA}$, they attained poor scores overall.

| | Index | $d_{QCD}$ | $d_W$ | $d_{GCC}$ | $d_J$ | $d_{VAR}$ | $d_{PCA}$ |
|---|---|---|---|---|---|---|---|
| Scenario 1 | ARI | 0.778 | 0.666 | 0.693 | **0.994** | 0.894 | 0.612 |
| | | (0.096) | (0.103) | (0.069) | (0.030) | (0.133) | (0.174) |
| | LA | 0.886 | 0.832 | 0.833 | **0.998** | 0.943 | 0.812 |
| | | (0.057) | (0.058) | (0.042) | (0.013) | (0.072) | (0.099) |
| | LOO1NN | 0.887 | 0.826 | 0.848 | **0.999** | 0.983 | 0.810 |
| | | (0.064) | (0.075) | (0.083) | (0.009) | (0.025) | (0.116) |
| | JI | 0.701 | 0.577 | 0.606 | **0.991** | 0.860 | 0.536 |
| | | (0.121) | (0.107) | (0.078) | (0.044) | (0.176) | (0.165) |
| Scenario 2 | ARI | **0.900** | 0.400 | 0.339 | 0.340 | 0.013 | 0.372 |
| | | (0.120) | (0.113) | (0.106) | (0.108) | (0.067) | (0.157) |
| | LA | **0.960** | 0.698 | 0.649 | 0.661 | 0.458 | 0.682 |
| | | (0.055) | (0.078) | (0.062) | (0.070) | (0.053) | (0.097) |
| | LOO1NN | **0.979** | 0.589 | 0.568 | 0.617 | 0.218 | 0.604 |
| | | (0.0378) | (0.109) | (0.130) | (0.118) | (0.109) | (0.147) |
| | JI | **0.866** | 0.374 | 0.344 | 0.336 | 0.152 | 0.368 |
| | | (0.146) | (0.085) | (0.065) | (0.072) | (0.035) | (0.113) |
| Scenario 3 | ARI | **0.893** | 0.046 | 0.003 | 0.401 | -0.009 | 0.009 |
| | | (0.137) | (0.123) | (0.084) | (0.133) | (0.077) | (0.100) |
| | LA | **0.963** | 0.547 | 0.522 | 0.722 | 0.509 | 0.523 |
| | | (0.051) | (0.075) | (0.065) | (0.066) | (0.057) | (0.067) |
| | LOO1NN | **0.973** | 0.415 | 0.328 | 0.587 | 0.297 | 0.299 |
| | | (0.043) | (0.138) | (0.144) | (0.137) | (0.134) | (0.166) |
| | JI | **0.874** | 0.213 | 0.193 | 0.425 | 0.195 | 0.192 |
| | | (0.158) | (0.067) | (0.048) | (0.088) | (0.043) | (0.054) |

Table 1: Averages and standard deviations (in brackets) of four clustering validity indexes for measures $d_{QCD}$, $d_W$, $d_{GCC}$, $d_J$, $d_{VAR}$ and $d_{PCA}$, according to the 100 trials of the simulation procedure. For each scenario and index, the best result is shown in bold. The length of each series was $T = 100$.

We repeated the simulations for Scenario 1 and Scenario 2, but this time the processes $(\epsilon_{t,1}, \epsilon_{t,2})'$ and $(\epsilon_{t,1}, \epsilon_{t,2}, \epsilon_{t,3})'$ were generated from a multivariate t distribution with 3 degrees of freedom. This allowed us to check the performance of the analysed dissimilarities under some amount of fat-tailedness in the error distribution. This feature is frequently exhibited by some series, mainly within the field of Finance. Therefore, it is reasonable to introduce fat-tailedness in the simulations, especially in Scenario 2, since dynamic conditional correlation models originally arose to model financial time series of stock returns.

The results involving this new distribution for the error terms are given in Table 2. Observing the latter, one can reach several conclusions. In Scenario 1, the quantile cross-spectral dissimilarity $d_{QCD}$ does not seem to be affected by the fat-tailedness of the error distribution, achieving a value of 0.772 for the ARI, versus 0.778 when the error terms follow a normal distribution (see Table 1). On the contrary, the remaining dissimilarities decreased their performance, especially $d_J$ and $d_{PCA}$. The model-based dissimilarity $d_{VAR}$ was the one achieving the best results, followed closely by the ones reached by $d_{QCD}$. In Scenario 2, all the dissimilarities worsened their performance, but $d_{QCD}$ was still the measure getting the highest scores. According to ARI, its average scores were at least twice as large as those reached by all the remaining dissimilarities.

The results obtained throughout this section are very powerful, since they indicate that the quantile cross-spectral dissimilarity not only offers the best general performance when a wide variety of situations are taken into account, but it is also quite robust to changes in the error distributions. This makes $d_{QCD}$ probably one of the best dissimilarities for practitioners to perform time series clustering, as it is well known that complex dependence patterns beyond linearity arise frequently in real MTS. Moreover, often normality of the error terms can not be guaranteed in practice.

## 4 A case study: Clustering bivariate series of daily returns and trading volume of some S&P 500 companies

In this section, the dissimilarity based on the quantile cross-spectral density, $d_{QCD}$, is used to perform clustering on a real data example involving financial time series. We consider both the daily stock prices and trading volume of some companies belonging to the S&P 500 index, which comprises 505 common stocks

| | Index | $d_{QCD}$ | $d_W$ | $d_{GCC}$ | $d_J$ | $d_{VAR}$ | $d_{PCA}$ |
|---|---|---|---|---|---|---|---|
| Scenario 1 | ARI | 0.772 | 0.577 | 0.625 | 0.650 | **0.845** | 0.262 |
| | | (0.092) | (0.083) | (0.093) | (0.111) | (0.140) | (0.124) |
| | LA | 0.883 | 0.766 | 0.797 | 0.806 | **0.917** | 0.589 |
| | | (0.057) | (0.058) | (0.055) | (0.074) | (0.077) | (0.088) |
| | LOO1NN | 0.884 | 0.763 | 0.785 | 0.914 | **0.970** | 0.519 |
| | | (0.059) | (0.093) | (0.092) | (0.056) | (0.034) | (0.127) |
| | JI | **0.692** | 0.495 | 0.540 | 0.568 | 0.650 | 0.261 |
| | | (0.111) | (0.073) | (0.085) | (0.116) | (0.111) | (0.078) |
| Scenario 2 | ARI | **0.775** | 0.359 | 0.282 | 0.352 | 0.000 | 0.313 |
| | | (0.116) | (0.116) | (0.098) | (0.115) | (0.070) | (0.121) |
| | LA | **0.897** | 0.671 | 0.627 | 0.667 | 0.452 | 0.646 |
| | | (0.067) | (0.075) | (0.064) | (0.068) | (0.056) | (0.073) |
| | LOO1NN | **0.898** | 0.552 | 0.490 | 0.527 | 0.222 | 0.480 |
| | | (0.074) | (0.114) | (0.135) | (0.109) | (0.115) | (0.172) |
| | JI | **0.713** | 0.352 | 0.305 | 0.348 | 0.144 | 0.316 |
| | | (0.134) | (0.077) | (0.061) | (0.080) | (0.037) | (0.081) |

Table 2: Averages and standard deviations (in brackets) of four clustering validity indexes for measures $d_{QCD}$, $d_W$, $d_{GCC}$, $d_J$, $d_{VAR}$ and $d_{PCA}$, according to the 100 trials of the simulation procedure. Innovations were drawn from a multivariate t distribution with 3 degrees of freedom. For each scenario and index, the best result is shown in bold. The length of each series was $T = 1000$.

issued by 500 large-cap companies and traded on American stock exchanges. The S&P 500 is commonly divided in eleven sectors. Only the companies belonging to financial and utilities sectors are considered. The reason of this choice rests on the fact that these two sectors are clearly different, in the sense that the activities performed by any two companies pertaining to each one of them greatly differ from one another. This is highly desirable, as our main purpose is to show to what extent $d_{QCD}$ is able to distinguish between series belonging to highly different economic sectors, supposedly representing strongly different economic behaviours. Had we chosen somehow overlapping sectors (e.g., all the eleven sectors) and the achieved conclusions could be misguided. The financial sector consists of banks, insurance companies, credit card issues and a host of other money-centric enterprises. The utilities sector includes many local electricity and water companies, among many others. Our database contains information about 55 companies belonging to the financial sector and 25 companies belonging to the utilities sector. The sample period spans from 8th February 2013 to 7th February 2018, thus resulting serial realizations of length $T = 1258$. The data are sourced from the Kaggle repository of Cameron Nugent[1] (Nugent, 2017).

Each of the 80 considered companies is then described by means of a bivariate time series whose components are the company's daily stock price and trading volume. The relationship between price and volume has been extensively analyzed in the literature (Karpoff, 1987; Campbell, Grossman, & Wang, 1993; Gebka & Wohar, 2013) and constitutes itself a topic of great financial interest. Prices and trading volume are known to exhibit some empirical linkages over the fluctuations of stock markets. Here, however, our concern is not to analyze whether these empirical facts hold true in the considered series, but assessing if the study of the joint behaviour of prices and volume via $d_{QCD}$ can give insights into the sector to which a company pertains. It seems reasonable to hypothesise that the joint behaviour of prices and volume shows some distinctive features depending on the sector. It can be observed that both the UTS of prices and trading volume are non-stationary in mean. For this reason, all UTS are transformed by taking the first differences of the natural logarithm of the original values. This way, prices give rise to stock returns, and volume, to what we call change in volume. This transformation is common when dealing with this kind of series (Chen, 2012). Finally, both series are normalized to have zero mean and unit variance. As an example, we depicted 9 of the series in the financial sector. They are shown in Fig. 1. In all cases, the blue color corresponds to the change in trading volume, while the red color corresponds to the stock returns. We have included the symbol of each one of the companies as given in the S&P 500.

Similar to other financial time series, stock returns exhibit empirical statistical regularities, so-called "stylized-facts". It is crucial to be aware of them in order to perform a proper analysis. The most common stylized facts include: heavy tails and a peak center compared to the normal distribution, volatility clustering (periods of low volatility mingle with periods of high volatility), leverage effects (returns are negatively correlated with volatility), and autocorrelation at much longer horizons than expected. In the same way, trading volume is known to empirically depart from normality. Table 3 provides some descriptive statistics regarding the returns and change in volume of the series of the top left panel of Fig. 1, which corresponds to the company Aflac. We can see that both the returns and the change in volume are skewed. In addition, the value of the kurtosis for the returns is 4.14, thus implying fatter tails than those of the normal distribution. Then, our proposal is to take advantage of the high capability of the quantile cross-spectral density to detect these stylized facts and performing cluster analysis based on $d_{QCD}$. In fact,

[1] `https://www.kaggle.com/camnugent/sandp500?`

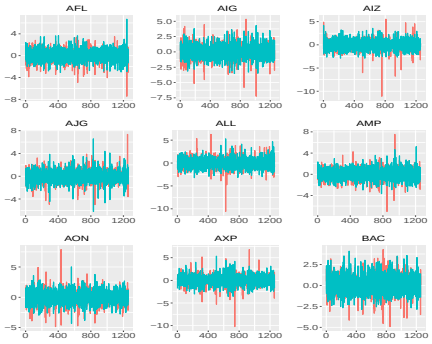



Figure 1: Bivariate series of returns (red color) and change in volume (blue color) for 9 companies of the S&P 500 belonging to the financial sector.

| Descriptive statistics | Returns | Change in volume |
|---|---|---|
| Minimum | -7.4529 | -3.1000 |
| Maximum | 3.5026 | 6.7142 |
| Skewness | -0.7321 | 0.4207 |
| Kurtosis | 4.1400 | 2.0883 |

Table 3: Descriptive statistics of returns and change in volume for the series in the top left panel of Fig. 1.

$d_{QCD}$ yielded by far the best average results classifying processes whose error terms exhibited some degree of fat-tailedness (see Table 2). Given its great performance in this kind of situations, we hope that, in the current scenario, $d_{QCD}$ can distinguish two clusters, one mostly formed by the companies in the financial sector and the other mainly constituted of the companies in the utilities sector.

The 80 bivariate series of returns and change in volume were subjected to PAM algorithm with the proposed dissimilarity, $d_{QCD}$. Just as in the simulations, $r = 3$ quantiles of levels 0.1, 0.5 and 0.9 along with the Fourier frequencies were considered to compute $d_{QCD}$. Once obtained the clustering solution, the performance of $d_{QCD}$ was assessed by comparing this solution with the assumed true partition, which is given by the sectors to which each company pertains. The same four indexes considered in Section 3 were used here to quantify the performance of the proposed metric.

In order to show to what extent $d_{QCD}$ is better able to detect the underlying structure in the data than other dissimilarities, the competitors in Section 3.2 were considered in the problem of grouping the financial time series. For all the dissimilarities, the same settings as in Section 3 were considered. It is worth noting that the selection of one lag for $d_{GCC}$ seems appropriate, given that the 1-lagged returns have been proven to have a highly predictive ability over the trading volume (Chen, 2012).

Table 4 shows the values achieved by the six dissimilarities with regards to each one of the indexes. It can be seen that $d_{QCD}$ got the best results, clearly outperforming the remaining dissimilarity measures in terms of all the indexes considered. Regarding the ARI, the proposed measure achieved 0.718, while its nearest competitor, $d_W$, only achieved 0.352. The generalized cross correlation-based distance and the PCA-based metric slightly detected some structure in the data, whereas the Maharaj's distance and the nonparametric dissimilarity $d_J$ were not capable of discriminating the series in both sectors at all. Similar insights can be obtained from the results with regards to the rest of the indexes. For instance, the high value of LOO1N that $d_{QCD}$ attained indicates that this metric is indeed the most appropriate to distinguish between underlying sectors in this kind of series.

In order to better understand the solution reached by $d_{QCD}$, Table 5 shows the number of companies

| Measure | ARI | LI | LOO1NN | JI |
|---|---|---|---|---|
| $d_{QCD}$ | **0.718** | **0.918** | **0.963** | **0.771** |
| $d_W$ | 0.352 | 0.790 | 0.863 | 0.532 |
| $d_{GCC}$ | 0.147 | 0.700 | 0.813 | 0.433 |
| $d_J$ | 0.030 | 0.641 | 0.863 | 0.567 |
| $d_{VAR}$ | 0.023 | 0.598 | 0.663 | 0.381 |
| $d_{PCA}$ | 0.167 | 0.616 | 0.725 | 0.574 |

Table 4: Four clustering validity indexes for measures $d_{QCD}$, $d_W$, $d_{GCC}$, $d_J$, $d_{VAR}$ and $d_{PCA}$, with regards to the financial time series.

| | No. companies in the financial sector | No. companies in the utilities sector |
|---|---|---|
| Cluster 1 | 2 | 24 |
| Cluster 2 | 53 | 1 |

Table 5: Breakdown of the number of companies in each sector located in each cluster, with regards to the clustering solution achieved by $d_{QCD}$.

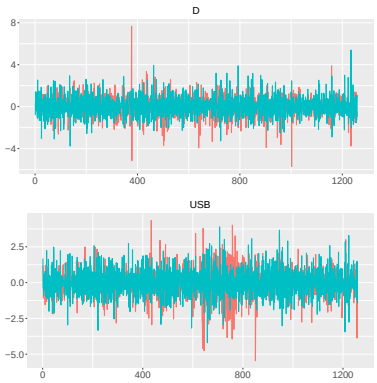



Figure 2: Top panel: bivariate series of returns (red color) and change in volume (blue color) of Dominion Energy (medoid of Cluster 1). Bottom panel: bivariate series of returns (red color) and change in volume (blue color) of U.S. Bancorp (medoid of Cluster 2).

pertaining to each sector which fell in each one of the clusters. It can be noticed that all but one of the companies in the utilities sector were located in the first cluster, along with 2 companies belonging to the financial sector. On the other hand, the second cluster contains the remaining 53 companies in the financial sector along with one company in the utilities sector.

The general impression that one gets from Table 5 is that the joint behaviour of price and volume is strongly related to the sector a given company pertains to, and that $d_{QCD}$ is indeed able to distinguish between the different underlying relationships of dependence. It seems that, in real-life situations like this, the approach followed by this metric, aimed to capture any type of dependence, gives it a pivotal advantage over the remaining measures, which are devised to uncover particular types of dependence as the linear one. By analysing this type of series via $d_{QCD}$, an investor could realise, for instance, that there are 2 particular companies in the financial sector whose behaviour is similar to that of the companies in the utilities sector, giving her valuable information about the market.

Perhaps one of the biggest advantages of using the PAM algorithm to perform clustering is that it offers a prototype for each encountered group. These prototypes are usually known as medoids, and they actually pertain to the original set that was subjected to the clustering procedure, giving them high interpretation. In this example, the medoids are bivariate time series which represent all the time series belonging to each cluster. They synthesize the cluster information and represent the prototype features of the clusters, then summarizing the characteristics of the time series within each group. Given this informative power of medoids, it is undeniably interesting to know which company is playing the role of prototype in both sectors.

The medoid time series of both sectors are depicted in Fig. 2. The top panel corresponds to the medoid bivariate time series within the cluster containing the companies in the utilities sector (Cluster 1). It represents the company Dominion Energy, commonly referred to as Dominion, a power and energy company headquartered in Richmond, which supplies electricity and natural gas in different parts of the United States. The bottom panel, displaying the medoid series regarding the financial cluster (Cluster 2), corresponds to U.S. Bancorp, a bank holding company based in Minneapolis, Minnesota, which provides banking, investment, and payment service products, among others. An investor could use the companies associated with both medoids as a proxy to help explain the financial situation of the corresponding sectors. Maybe, analysing these two companies gives more valuable insights into the future of both sectors than performing an extensive study over the whole groups.

It is difficult to work out from Fig. 2 that those two series correspond to two kinds of financial behaviours which are profoundly different. This is in part due to the length of both series, 1258, which makes difficult for the eye to detect any pattern. Besides, this is probably also attributable to the complex forms of dependence that exist between the price and the volume of each company. For instance, we have

seen in Table 4 that $d_{GCC}$, a dissimilarity measure based on an intuitive quantity as the cross-correlation, barely noticed the existence of two different underlying sectors.

Given the previous results, we are compelled to emphasize that practitioners in the field of MTS clustering should take into account the dissimilarity based on the quantile cross-spectral density, capable of detecting patterns that could remain invisible otherwise.

## References

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, *53*, 16–38.

Alonso, A. M., & Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing*, *29*(4), 655–676.

Baruník, J., & Kley, T. (2019). Quantile coherency: A general measure for dependence between cyclical economic variables. *The Econometrics Journal*, *22*(2), 131-152.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993). Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, *108*(4), 905–939.

Chen, S.-S. (2012). Revisiting the empirical linkages between stock returns and trading volume. *Journal of Banking & Finance*, *36*(6), 1781–1788.

Dette, H., Hallin, M., Kley, T., & Volgushev, S. (2015, 05). Of copulas, quantiles, ranks and spectra: An $l_1$-approach to spectral analysis. *Bernoulli*, *21*(2), 781–831.

D'Urso, P., & Maharaj, E. A. (2012). Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems*, *193*, 33–61.

Fu, T.-c. (2011, February). A review on time series data mining. *Eng. Appl. Artif. Intell.*, *24*(1), 164–181.

Gebka, B., & Wohar, M. E. (2013). Causality between trading volume and returns: Evidence from quantile regressions. *International Review of Economics & Finance*, *27*, 144–159.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Kakizawa, Y., Shumway, R. H., & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, *93*(441), 328–340.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and quantitative Analysis*, 109–126.

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, *7*(4), 349–371.

Kley, T. (2016). Quantile-based spectral analysis in an object-oriented framework and a reference implementation in R: The **quantspec** package. *Journal of Statistical Software*, *70*(3), 1–27. doi: 10.18637/jss.v070.i03

Kley, T., Volgushev, S., Dette, H., & Hallin, M. (2016, 08). Quantile spectral processes: Asymptotic analysis and inference. *Bernoulli*, *22*(3), 1770–1807.

Lafuente-Rego, B., & Vilar, J. A. (2016). Clustering of time series using quantile autocovariances. *Advances in Data Analysis and classification*, *10*(3), 391–415.

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining* (pp. 16–22).

Lee, J., & Rao, S. S. (2012). *The quantile spectral density and comparison based tests for nonlinear time series.*

Liao, T. W. (2005). Clustering of time series data: A survey. *Pattern Recognit.*, *38*(11), 1857–1874.

López-Oriona, Á., Vilar, J. A., et al. (2021). Quantile-based fuzzy clustering of multivariate time series in the frequency domain. *arXiv preprint arXiv:2109.03728*.

Maharaj, E. (1999). Comparison and classification of stationary multivariate time series. *Pattern Recognition*, *32*(7), 1129–1138.

Maharaj, E., D'Urso, P., & Caiado, J. (2019). *Time series clustering and classification.* CRC Press.

Nugent, C. (2017). S&p500 stock data. Kaggle. `https://www.kaggle.com/camnugent/sandp500?` (Accesed 15 July 2020)

Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J., & Keogh, E. (2017). Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, *31*(1), 1–31.

Singhal, A., & Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *19*(8), 427–438.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# CÁLCULO DE ÍNDICES DE PODER PARA JUEGOS DE MAYORÍA PONDERADA

Livino M. Armijos-Toro[1,2], José María Alonso-Meijide[1] y Manuel A. Mosquera[3]

[1]Departamento de Estadística, Análisis Matemático y Optimización. Universidad de Santiago de Compostela. España.
[2]Departamento de Ciencias Exactas. Universidad de las Fuerzas Armadas ESPE. Ecuador.
[3]Departamento de Estadística e Investigación Operativa. Universidad de Vigo. España.

## RESUMEN

El cálculo de índices de poder para juegos de mayoría ponderada puede ser realizado mediante diferentes procedimientos. Uno de estos son las funciones generatrices (Wilf, 1994; Alonso-Meijide et al., 2009; Alonso-Meijide et al., 2012). En este trabajo, se proponen métodos de cálculo para los índices de poder de Johnston (Johnston, 1978) y Colomer-Martínez (Colomer & Martínez, 1995) mediante el uso de funciones generatrices. Además, nosotros proponemos un nuevo índice de poder para juegos de mayoría ponderada que combina los índices de poder de Johnston y Colomer-Martínez. Se propone un método de cálculo de este nuevo índice de poder mediante funciones generatrices.

**Palabras y frases clave**: juegos simples, juegos de mayoría ponderada, índices de poder, funciones generatrices.

## REFERENCIAS

Alonso-Meijide, J.M. and Bilbao, J. and Casas-Méndez, B. and Fernández, J. (2009) Weighted multiple majority games with unions: Generating functions and applications to the European Union. European Journal of Operational Research. V 198 - N 2, 530 – 544.

Alonso-Meijide, J.M. and Freixas, J. and Molinero, X. (2012) Computation of several power indices by generating functions. Applied Mathematics and Computation. V 219 - N 8, 3395 – 3402.

Colomer, J.M. and Martínez, F. (1995) The paradox of coalition trading. Journal of Theoretical Politics. V 7 - N 1, 41 – 63.

Johnston, R.J. (1978) On the measurement of power: Some reactions to Laver. Environment and Planning A: Economy and Space. V 10 - N 8, 907 – 914.

Wilf, H. (1994) Generatingfunctionology (Second Edition). Academic Press.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# PAIRWISE JUSTIFIABLE CHANGES IN COLLECTIVE CHOICES

Salvador Barberá[1], Dolors Berga[2], Bernardo Moreno[3] and Antonio Nicolò[4]

[1]MOVE, Universitat Autònoma de Barcelona, and Barcelona GSE. Mailing address: Departament d'Economia i d'Història Econòmica, Edifici B, 08193 Bellaterra, Spain. E-mail: salvador.barbera@uab.cat

[2]Departament d'Economia, C/ Universitat de Girona, 10; Universitat de Girona, 17071 Girona, Spain. E-mail: dolors.berga@udg.edu

[3]Departamento de Teoría e Historia Económica, Facultad de Ciencias Económicas y Empresariales, Campus de El Ejido, 29071 Málaga, Spain. E-mail: bernardo@uma.es

[4]University of Padova, via del Santo, 33 - 35122 Padova, Italy. E-mail: anto- nio.nicolo@unipd.it; University of Manchester, Arthur Lewis Building, School of Sciences, Economics, Manchester, UK.

## ABSTRACT

Consider the following principle regarding the performance of collective choice rules. "If a rule selects alternative x in situation 1, and alternative y in situation 2, there must be an alternative z, and some member of society whose appreciation of z relative to x has increased when going from situation 1 to situation 2." This principle requires a minimal justification for the fall of x in the consideration of society: someone must have decreased its appreciation relative to some other possible alternative. We study the consequences of imposing this requirement of pairwise justifiability on a large class of collective choice rules that includes social choice and social welfare functions as particular cases. When preference profiles are unrestricted, it implies dictatorship, and both Arrow's and the Gibbard-Satterthwaite theorems become corollaries of our general result. On appropriately restricted domains, pairwise justifiability, along with anonymity and neutrality, characterize Condorcet consistent rules, thus providing a foundation for the choice of the alternatives that win by majority over all others in pairwise comparisons, when they exist.

**Keywords**: Pairwise justifiability, preference reversal condition, collective choice correspondences, Condorcet consistency, Arrow's theorem, Gibbard-Satterthwaite's theorem.

## REFERENCES

Arrow, K. (1963). *Social Choice and Individual Values*. 2nd edition New York: Wiley (1st edition 1951).

Eliaz, K. (2004). Social Aggregators. *Social Choice and Welfare*, Vol. 22, 2, 317-330.

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41: 587-601.

Satterthwaite, M. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41,4: 587-601.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ON THE COOPERATION IN SEQUENCING SITUATIONS WITH EXPONENTIAL POSITIONAL EFFECTS

Alejandro Saavedra-Nieves[1], Manuel A. Mosquera[2] and M. Gloria Fiestras-Janeiro[2]

[1]Universidade de Santiago de Compostela, Departamento de Estatística, Análise Matemática e Optimización.
[2]Universidade de Vigo, Departamento de Estatística e Investigación Operativa.

## ABSTRACT

Sequencing problems describe situations where several jobs have to be processed on a set of machines. This class of problems are formally characterized by an initial order for the jobs and a family of cost functions by their processing. In this sense, different factors, as the starting time or the position in the queue of a job, naturally affect its real processing time.

Cooperation in sequencing problems was widely treated in literature with the aim of minimizing the total costs of processing. In this sense, the usual idea of savings in sequencing can be naturally extended to any other setting in which the considered optimality criteria is not constant over all orders. Under positional effects, other classical measures, as the makespan, may be considered since that, in this setting, savings also come from repairing positions of jobs. However, two common issues have to be still addressed: identify the optimal sequence for the jobs, and distribute the associated savings of the measure of interest with respect to the initial order among the agents using cooperative game theory.

In this work, we analyse sequencing problems with exponential positional effects of learning and deterioration of the machine. Specifically, we obtain some results on the optimal order for the makespan and for the total flow time and analyse the cooperation in sequencing through the saving games associated to these situations.

**Keywords**: exponential positional effects, position-dependent neighbour switching gains, learning and deterioration, convexity, stable allocations.

## REFERENCES

Borm, P., Fiestras-Janeiro, G., Hamers, H., Sánchez, E., Voorneveld, M. (2002). On the convexity of games corresponding to sequencing situations with due dates. European Journal of Operational Research, 136(3), 616-634.

Curiel, I., Pederzoli, G., Tijs, S. (1989). Sequencing games. European Journal of Operational Research, 40(3), 344-351.

Gordon, V. S., Potts, C. N., Strusevich, V. A., Whitehead, J. D. (2008). Single machine scheduling models with deterioration and learning: handling precedence constraints via priority generation. Journal of Scheduling, 11(5), 357–370.

Saavedra-Nieves, A., Schouten, J., Borm, P. (2020). On interactive sequencing situations with exponential cost functions. European Journal of Operational Research, 280(1), 78-89.

Schouten, J., Saavedra-Nieves, A., Fiestras-Janeiro, M. G. (2021). Sequencing situations and games with non-linear cost functions under optimal order consistency. European Journal of Operational Research, 294(2), 734-745

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# A NEW ALGORITHM FOR INFLUENCE MAXIMIZATION

Elisenda Molina[1], Juan Tejada[2] and Juan Vidal-Puga[3]

[1] Universidad Carlos III de Madrid
[2] Universidad Complutense de Madrid
[3] Universidade de Vigo

## ABSTRACT

We present a stochastic discounted heuristic algorithm for the maximization problem where the aim is to find the most influential nodes in a network. This algorithm is efficient and faster than other heuristics presented in the literature in all of the checked networks.

**Keywords:** Influence maximization, network, algorithm, heuristic.

## 1. INTRODUCTION

In this paper, we study dissemination of innovations, habits, rumours, propaganda, social movements, strikes, product consumption, fashions, cascading failures (physical, financial systems), etc. through a network when the propagation of the information is a result of the influence that one individual can exert over other.

Applications cover many different fields, such as marketing, sociology and social psychology, epidemiology, criminology, system reliability, etc.

In particular, we address the problem of finding a subset of $k$ seed nodes for starting the diffusion process from them (Granovetter (1978), Domingos and Richardson (2001), Kempe *et al.* (2003)).

The key elements are the following:

- A Social network (SN), $(V, \Gamma)$, and a Socio-matrix, $W = (w_{ij})$ such that $w_{ij}$ represents the influence that agent $i$ exerts over agent $j$.
- A diffusion model.

## 2. THE MODEL

In this section, we describe the diffusion information models. Formally, a (stochastic) diffusion information model $\{\mathbf{X}(t)\}_{t\geq 0}$ is defined by $X_v(t) \in \{0,1\}$, $v \in V$, governed by the following *local interaction* rules:

1. We do not consider strategic aspects in the interaction between individuals.
2. The probability that an individual adopts the innovation depends on their degree of exposure to such innovation in the social network and on how this exposure affects her.
3. *Progressive*: Once the innovation has been adopted, its abandonment is no longer considered.

The most common diffusion information models is the following:

**ICM. Independent Cascade Model (Kempe et al., 2003)**

- Socio-matrix, $\mathbf{W}$: $w_{uv}$ := influence that agent $u$ exerts over agent $v$.
  - Followers of $u \in V$: $N^+(u) := \{v \in V : w_{uv} > 0\}$.
  - Influencers of $u \in V$: $N^-(u) := \{v \in V : w_{vu} > 0\}$.
- A seed group of members $T \subseteq V$ who initially promote the innovation: $\mathbf{X}(0) := (\mathbf{0}_{V\setminus T}, \mathbf{1}_T)$.
- Transition probabilities:
  - Each active agent $u$ has a unique attempt to convince her non-active followers $v \in N^+(u)$ to adopt the innovation.

- Probability of convincing her: $w_{uv}$
- Attempts are made independently.
- Each active agent remains active.

- *Expected diffusion* of $T$, $\sigma_{LT}(T) = E_W[|T_F|]$, $T_F$ = set of active nodes in the final absorbing state.

Formarly, the ICM is defined as follows (weights should be 1 or less):
Input: $S$(set of nodes)

1. initialize infected(1...$n$) ← false
2. initialize ninfected ← 0
3. **if S = ∅, then return infected and ninfected**
4. for each $i$ in $S$, do
5. infected($i$) ← true
6. ninfected ← ninfected + 1
7. end for
8. initialize $T \leftarrow \emptyset$
9. while $S \neq \emptyset$, do
10. for each $i$ in $S$, do
11. for each $j$ pointing to $i$, do
12. if infected($j$) = false, then
13. if $w_{ji}$ > random number between 0 and 1, then
14. infected($j$) ← true
15. ninfected ← ninfected + 1
16. $T \leftarrow T \cup \{j\}$
17. end if
18. end if
19. end for
20. end for
21. $S \leftarrow T$
22. end while
23. return infected and ninfected

**IMP. Influence Maximization Problem**

Stochastic combinatorial optimization problem: find a *target set* $T^* \subseteq V$ with $|T^*| = k$ maximizing its expected influence under the considered stochastic diffusion model:

$$\max\{\sigma(T) : T \subseteq V, |T| = k\}.$$

The characteristics of the problem are the following:

- NP-hard problem.
- Expected diffusion function properties:
  - Monotonic (weakly increasing): $\sigma(S) \leq \sigma(T)$, for all $S \subseteq T \subseteq V$.
  - Submodular: $\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T)$ for all $S \subseteq T \subseteq V$.
- Greedy approach efficiency (Nemhauser *et al.*, 1978)

$$f(T^g) \geq (1 - 1/e)\, f(T^*)$$

The greedy method is as follows:

Input: $R$ (number of repetitions), $k$.

1. initialize $A \leftarrow \emptyset$
2. initialize $best_S \leftarrow 0$
3. for $m$ = 1 to $k$, do
4. initialize $best_i \leftarrow 0$
5. for each $i$ in $N$, do
6. initialize $B \leftarrow A \cup \{i\}$
7. initialize $s_B \leftarrow 0$
8. repeat $R$ times
9. assign random thresholds to nodes
10. $s_B \leftarrow s_B + \sigma(S)$
11. end repeat

12. $s_B \leftarrow s_B / R$
13. if $s_B > \text{best}_S$, do
14. $\text{best}_S \leftarrow s_B$
15. $\text{best}_i \leftarrow i$
16. end if
17. end for
18. $A \leftarrow A \cup \{\text{best}_i\}$
19. end for
20. Declare the nodes in $A$ as top-$k$ nodes

In the IMP, the greedy approach algorithms have several drawbacks due to no scalability. The greedy algorithm is based on intensive Montecarlo estimations ($\sigma(S)$ must be approximated). Some improvements include the estimation of influence spread by means of simulating the *live-arc graph* and reducing the number of marginal contributions to be computed, as in the CELF algorithm (Leskovec *et al.*, 2007). Scalable alternatives are based on alternative estimations of individual spreading ($\sigma(\{v\})$), marginal contributions ($MC(v,\sigma(T) = \sigma(T \cup \{v\}) - \sigma(T)$), and by means of influence paths, reverse reachable sets, etc.

## 2. HEURISTICS

The Cost Effective Lazy Forward (CELF) method, proposed by Leskovec *et al.* (2007), provides the same results as the greedy algorithm but it is much faster under submodularity.
Initialize $A \leftarrow \emptyset$

1. for each $i$ in N, do
2. $\delta_i \leftarrow \infty$
3. end for
4. repeat $k$ times
5. for each $i$ in $N \setminus A$, do
6. $\text{cur}_i \leftarrow$ false
7. end for
8. repeat forever
9. initialize $\Delta = 0$, $i^*$
10. for each $i$ in $N \setminus A$, do
11. if $\delta_i > \Delta$, then
12. $\Delta = \delta_i$
13. $i^* = i$
14. end if
15. end for
16. if $\text{cur}_{i^*}$, then
17. $A \leftarrow A \cup \{i^*\}$
18. break
19. else
20. $\delta_{i^*} \leftarrow \sigma(A \cup \{i^*\}) - \sigma(A)$
21. $\text{cur}_{i^*} \leftarrow$ true
22. end if
23. end repeat
24. end repeat
25. Declare the nodes in $A$ as top-$k$ nodes

The Directed Degree Heuristic (DDH, Chen et al., 2019) is a constructive algorithm based on assuming a homogeneous IC model ($w_{uv} = w_{vu} = p$ for each pair $u, v \in V$) and local influence (only her direct followers) to approximate $\sigma(\{i\})$ and $MC(v,\sigma(T))$ disregarding indirect effects.

Step 1: Select the agent with maximum degree $d_v$.

Step $t + 1$: Expected *Star Marginal Contribution* approximation:

$$SMC_v \approx (1-p)^{t_v} \left(\Sigma_{u \in N(v) \setminus T}\, p + 1\right) \approx 1 + (d_v - 2t_v - (d_v - t_v)\, t_v p)\, p \;,\; v \in V \setminus T,$$

where $T$ is the current set of selected agents and $t\_v = t_v = |T(v)| = |N(v) \cap T|$ is the number of her directed contacts already in $T$.

We then select the agent with maximum *discounted degree*

$$dd_v := d_v - 2t_v - (d_v - t_v)t_v p.$$

The DDH method is as follows:

Initialize $S \leftarrow \emptyset$

1. for each vertex $v$ do
2. compute its degree $d_v$
3. $dd_v \leftarrow d_v$
4. initialize $t_v$ to 0
5. end for
6. for $i = 1$ to $k$ do
7. select $u = \arg\max_v\{dd_v : v \in V \setminus S\}$
8. $S \leftarrow S \cup \{u\}$
9. for each neighbor $v$ of $u$ and $v \in V \setminus S$ do
10. $t_v \leftarrow t_v + 1$
11. $dd_v \leftarrow d_v - 2t_v - (d_v - t_v)t_v p$
12. end for
13. end for
14. output $S$

Chen *et al.* (2009) use $p = 0.01$.

## 2. THE STOCHASTIC DISCOUNTED HEURISTIC

For each i, j ∈ N, let $p_{ij}$ be the probability that j is contaminated when i is the only initial node. Hence, after R repetitions, we estimate $p_{ij}$ as the number of times that j is contaminated in $\sigma(\{i\})$ over $R$. Clearly, the probability of node $j$ be contaminated is not independent from the probability that some other node (say, $l$) be contaminated also. However, it is not difficult to check that , es fácil comprobar que $\sigma(\{i\})) = \sum_{j\in N} p_{ij}$. The first step in our algorithm is to take $i_1 \in \operatorname{argmax}_{i\in N} \sum_{j\in N} p_{ij}$. This step coincides with greedy and CELF.

To choose the second node, we assume that the probabilities $p_{ij}$ are still approximately the same as before, because we are interested in finding nodes far away from the influence of the first one, $i_1$. A node $j$ will be contaminated in this second stage with probability $p_{i_1 j}$, and hence we focus on how to contaminate the remaining $1 - p_{i_1 j}$.

Thus, the second step is to take $i_2 \in \operatorname{argmax}_{i\in N\setminus\{i1\}} \sum_{j\in N} p_{ij}(1 - p_{i_1 j})$.

In general, in step $k$ we take $i_k \in \operatorname{argmax}_{i\in N\setminus\{i1,\ldots,ik\text{-}1\}} \sum_{j\in N} p_{ij} \prod_{m=1,\ldots,k\text{-}1}(1 - p_{imj})$.

However, this methods fails in some cases, due to the fact, by choosing some $i_m$, it may be already chosen with a high probability by previous nodes. This problem is solved by weighting not only by $1 - p_{i_1 j}$, but also by $1 - p_{i_1 i}$. Thus, the algorithm mis the following. In step $k$, we choose

$$i_k \in \operatorname{argmax}_{i\in N\setminus\{i1,\ldots,ik-1\}} \sum_{j\in N} p_{ij} \prod_{m=1,\ldots,k-1}(1 - p_{imj})(1 - p_{imi}).$$

Formally:

1. Initialize $p[1...n][1...n]$, $q[1...n][1...n]$ to 0,
2. Repeat $R$ times
3. assign random thresholds to nodes
4. for each $i$ in $N$, do
5. compute $\Sigma(i)$, the set of nodes infected by $i$
6. for each $j$ in $\Sigma(i)$, do
7. $p[i][j] \leftarrow p[i][j] + 1/R$
8. $q[i][j] \leftarrow p[i][j]$
9. end for
10. end for

11. end repeat
12. initialize A ← ∅, lastknode
13. for $m$ = 1 to $k$, do
14. initialize maxsum ← 0
15. for each $i$ in $N \setminus A$, do
16. initialize sum ← 0
17. for each $j$ in $N$, do
18. if $m > 1$, then
19. $q[i][j] \leftarrow q[i][j] * (1 - p[\text{lastknode}][j]) * (1 - p[\text{lastknode}][i])$
20. end if
21. sum ← sum + $q[i][j]$
22. end for
23. if sum > maxsum, then
24. maxsum ← sum
25. lastknode ←$i$
26. end if
27. end for
28. $A \leftarrow A \cup \{\text{lastknode}\}$
29. end for
30. present $A$ as top-$k$ nodes

The next modification is also scalable:

1. For each $i$ in $N$, initialize $S_i \leftarrow \emptyset$
2. Repeat $R$ times
3. assign random thresholds to nodes
4. for each $i$ in $N$, do
5. compute $\Sigma(i)$, the set of nodes infected by $i$
6. for each $j$ in $\Sigma(i)$, do
7. if $j$ in $S_i$, then
8. $p[i][j] \leftarrow p[i][j] + 1/R$
9. $q[i][j] \leftarrow p[i][j]$
10. else
11. $S_i \leftarrow S_i \cup \{j\}$
12. initialize $p[i][j] \leftarrow 0$, $q[i][j] \leftarrow 0$
13. end if
14. end for
15. end for
16. end repeat
17. initialize A ← ∅, lastknode
18. for $m$ = 1 to $k$, do
19. initialize maxsum ← 0
20. for each $i$ in $N \setminus A$, do
21. if $i$ in $S_{\text{lastknode}}$, then
22. define $w_i \leftarrow (1 - p[\text{lastknode}][i])$
23. else
24. define $w_i \leftarrow 1$
25. end if
26. initialize sum ← 0
27. for each $j$ in $S_i$, do
28. if $m > 1$, then
29. if $j$ in $S_{\text{lastknode}}$, then
30. define $w_j \leftarrow (1 - p[\text{lastknode}][j])$
31. else
32. define $w_j \leftarrow 1$
33. end if
34. $q[i][j] \leftarrow q[i][j] * w_i * w_j$
35. end if
36. sum ← sum + $q[i][j]$

```
37.       end for
38.       if sum > maxsum, then
39.          maxsum ← sum
40.          lastknode ← i
41.       end if
42.    end for
43.    A ← A ∪ {lastknode}
44. end for
45. present A as top-k nodes
```

## 3. RESULTS

The stochastic discounted heuristic presented in this paper provides an efficient algorithm, faster algorithm than other heuristics presented in the literature in all of the checked networks, including NetHEPT (Berman and Parikh, 2000), Epinions (Richardson and Agrawal, 2003), Polblogs, and p2pGnutella31 (Leskovec et at., 2007 and Ripeanu et al., 2002).

## REFERENCES

Berman, D. S. and Parikh, M. K. (2000). Confinement and the AdS/CFT Correspondence. Phys.Lett. B483, 271-276.

Chen W., Wang Y. and Yang S. (2009) Efficient influence maximization in social networks. KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 937-944.

Domingos, D. and Richardson, M. (2001). Mining the network value of customers. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 57-66.

Granovetter, M. (1978). Threshold Models of Collective Behavior. American Journal of Sociology 83, 489-515.

Kempe D., Kleinberg J. M., and Tardos É. (2003). Maximizing the spread of influence through a social network. Proceedings 9th ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD), 137-146.

Leskovec, J., Kleinberg, J. and Faloutsos, C. (2007). Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1).

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming 14, 265-294.

Richardson, M., Agrawal, R. and Domingos P. (2003). Trust Management for the Semantic Web. ISWC.

Ripeanu, M., Foster, I., and Iamnitchi A. (2002). Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. IEEE Internet Computing Journal.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# TÉCNICAS DE REGULARIZACIÓN ROBUSTAS PARA EL MODELO DE REGRESIÓN LOGÍSTICO.

Ghosh, A.[1]; Jaenada, M.[2] y Pardo, L.[2]

[1]Indian Statistical Institute, India
[2]Complutense University of Madrid, Spain

## RESUMEN

Los métodos clásicos de estimación parámetrica en modelos de regresión no son adecuados cuando el número de variables explicativas en el modelo es muy superior al número de observaciones. Las técnicas de regularización presentan una alternativa para efectuar simultáneamente selección de variables y estimación paramétrica. Además, es conocida la falta de robustez de los estimadores basados en la habitual función de verosimilitud, por lo que deben desarrollarse estimadores robustos que no se vean influenciados por contaminación en la muestra, mientras que se comporten adecuadamente en escenarios sin contaminación.

En este trabajo se desarrollan estimadores robustos basados en la conocida divergencia de mínima potencia para modelo de regresión logístico con datos de alta dimensión. Estos estimadores son demostrablemente robustos, y poseen convenientes propiedades asíntoticas. El modelo de regresión logística es herramienta apropiada para problemas clasificación, con una interpretación probabilística directa. Asimismo, el problema de clasificación en escenarios de alta dimensionalidad tiene una amplia aplicabilidad práctica, como se ilustra más adelante con el diágnostico de pacientes con cáncer basado en la expresión genética del paciente.

**Palabras y frases clave**: Divergencia de densidad de potencia, datos de alta dimensión, modelo de regresión logística, selección de variables, robustez.

## 1. INTRODUCCIÓN

El avance en la tecnología de recogida de datos y la capacidad de computación ha supuesto el desarrollo de numerosos modelos estadísticos adaptados a distintas situaciones. En particular, se han desarrollado técnicas de inferencia para los denominados datos de alta dimensión. Decimos que un conjunto de datos es de alta dimensión cuando el número de variables disponibles es muy superior al número de observaciones, pudiendo ser ésta una relación exponencial. Esta situación se presenta conmumente en distintas áreas del conocimiento, en las que se destaca la biomedicina y biogenética.

Los modelos de regresión relacionan una variable respuesta con un conjunto de variables explicativas mediante una transformación adecuada del predictor lineal. En particular, el modelo logístico modeliza la distribución de una respuesta dicotómica $Y$ según un vector de variables explicativas o covariables. El modelo de regresión logística ha sido ampliamente usado como clasificador, con una interpretación probabilistica directa. Sin embargo, las técnicas clásicas de estimación en el modelo logístico asumen que el número de observaciones $n$ es superior al número de covariables $p$, por lo que no son adecuadas en el contexto de datos de alta dimensión. A este respecto, se desarrollan las técnicas de regularización que llevan a cabo simultáneamente la estimación parámetrica y selección de variables.

Por último, es conocida la falta de robustez del clásico estimador de máxima verosimiltud. Esta sensibilidad a datos atípicos es aún más apreciable ante datos de alta dimensión, donde el número de observaciones es escaso y una observación anómala tiene mayor peso en la muestra.

Así, es necesario el desarrollo de técnicas de estimación robustas para datos de alta dimensión. En Ghosh y Basu (2016) se proponen una familia de estimadores robustos basados en la divergencia de densidad de potencia (DPD) para modelos lineales generalizados, particularizando en el modelo logístico para datos de baja dimensión. En Ghosh et al. (2021) se amplía la definición para datos de alta dimensión en el caso del modelo de regresión lineal, y se propone la extensión al modelo logístico como siguiente paso natural.

## 2. ESTIMADOR DE MÍNIMA DIVERGENCIA DE DENSIDAD DE POTENCIA PARA EL MODELO DE REGRESIÓN LOGÍSTICA

Sean $Y_1, ..Y_n$ un conjunto de variables aleatorias dicotómicas e independientes entre sí, siguiendo una distribución de Bernouilli,

$$P(Y_i = 1) = \pi_i, i = 1, .., n,$$

con $\pi_i \in [0, 1]$. El modelo logístico relaciona la probabilidades $\pi$ de la distribución de Bernouilli con un vector $p$-dimensional de variables explicativas, $\boldsymbol{x}_i$, a través de una función de enlace y el predictor lineal según,

$$\text{logit}(\pi_i) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} i = 1, .., n,$$

donde la función de enlace es $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ y el vector $\boldsymbol{\beta} = (\beta_1, .., \beta_p) \in \mathbb{R}^p$ es común para todas las observaciones.

La familia de DPD fue introducida por Basu et al. (1998), y Ghosh y Basu (2016) adaptaron la definición para modelos de regresión generalizados, obteniendo resultados teóricos de robustez y consistencia. En particular, dado un conjunto de datos $(y_1, \boldsymbol{x}_1), .., (y_n, \boldsymbol{x}_n)$, la función de pérdida basada en la DPD es de la forma

$$\begin{aligned} d_\alpha(\boldsymbol{\beta}) = \frac{1}{n^{1+\alpha}} \sum_{i=1}^{n} \Bigg\{ & \left(\pi^{1+\alpha}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \left(1 - \pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)^{1+\alpha}\right) \\ & - \left(1 + \frac{1}{\alpha}\right)\left(y_i \pi^\alpha(\boldsymbol{x}_i^T\boldsymbol{\beta}) + (1 - y_i)\left(1 - \pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)^\alpha\right) + \frac{1}{\alpha}\left(y_i^{\alpha+1} + (1 - y_i)^{\alpha+1}\right)\Bigg\}. \end{aligned} \quad (1)$$

donde el parámetro $\alpha \geq 0$ controla el equilibrio entre eficiencia y robusted de la función de pérdida. Como es habitual, se define el estimador de mínima DPD (MDPDE) como

$$\widehat{\beta}_\alpha = \min_{\beta \in \mathbb{R}} d_\alpha(\boldsymbol{\beta})$$

Nótese que la definión anterior no es válida para el valor $\alpha = 0$, pero pueden tomarse límites continuos, $\alpha \to 0$ para obtener la función de pérdida correspondiente. En este caso, el MDPDE obtenido coincide con el EMV.

Las técnicas de regularización introducen una función de penalización en el valor absoluto de los coeficientes del vector $\boldsymbol{\beta}$, $p_\lambda$, induciendo la estimación nula de muchos coeficientes y por tanto la selección de variables. Así, la función objetivo a minimizar para obtener un MDPDE regularizado, que también denotamos por MDPDE es de la forma

$$Q_\alpha(\boldsymbol{\beta}) = d_\alpha(\boldsymbol{\beta}) + p_\lambda(|\beta_j|)$$

donde el parámetro $\lambda$ controla el peso de la penalización, y por tanto la escasez de coeficientes no nulos en el vector estimado. Para la función de pérdida se han planteado distintas propuestas, entre las que destacamos la penalización LASSO introducida por Tibshirani (1996), y la penalización LASSO Adaptativo (Zou (2006)), una versión mejorada de la primera basada en una estimación inicial del vector de coeficientes.

Los métodos de optimización para la estimación de los parámetros en el modelo logístico regularizado con pérdida cuadrática emplean el algoritmo iterativo IRLS (*iteratively reweighted least squares* en inglés) que en cada paso resuelve un problema de regresión lineal auxiliar. Este método debe adaptarse al la pérdida basada DPD, a fin de obtener un algoritmo computacional válido. Además, se debe establecer un criterio para la selección del parámetro $\lambda$ óptimo. En nuestro trabajo, proponemos el criterio de información generalizada (GIC) adaptado a datos de alta dimensión (HGIC)

$$\text{HGIC}(\lambda) = \frac{-2\log\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda)}{n} + \frac{\log\log(n)\log(p)}{n}||\widehat{\boldsymbol{\beta}}||_0 \quad (2)$$

donde $\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda)$ denota la función de verosimilitud del modelo. Se elige $\lambda$ aquel valor con menor HGIC entre un mallado de valores predefinido.

## 3. APLICACIÓN A DATOS REALES

El modelo de regresión logístico para datos de alta dimensión puede ser útil en multitud de áreas del conocimiento, entre la que destacamos la genética. Para ilustrar la aplicabilidad del modelo propuesto, se utiliza una base de datos para la clasificación de pacientes con cáncer de pecho. El modelo de regresión pretende relacionar el diagnóstico con la expresión genética de los pacientes, pudiendo además identificar los genes relacionados con el desarrollo de la enfermedad. Así, el modelo persigue un doble objetivo, identificación de los genes implicados en la enfermedad y diágnostico del paciente.

La expresión genética de un individuo puede obtenerse mediante la tecnología de *microarray* recientemente desarrollada. Los datos de *microarrays* son generalmente datos de alta dimensión, con un gran número de genes en comparación con el número de muestras. Además, es bien sabido que los conjuntos de datos de *microarrays* con muchos genes suelen contener valores atípicos y varios estudios han señalado que los errores en el etiquetado y la medida de la expresión génica no son infrecuentes, lo que motiva el uso de procedimientos robustos.

## 3. CONCLUSIONES

El modelo de regresión logístico regularizado es una herramienta últil para clasificación y selección de variables de forma simultánea. Sin embargo, a pesar de ser el más eficiente, el EMV es poco robusto, lo que motiva el desarrollo de nuevos estimadores que se sean menos influenciados por contaminación en la muestra. Los MDPDE son demostrablemente robustos y se comportan adecuadamente ante datos puros, lo que hace esta propuesta muy atractiva. Además, el modelo logístico regularizado es aplicable muchos problemas reales, como es la identificación de genes involucrados en una determinada enfermedad y el diagnóstico de pacientes.

## REFERENCIAS

Basu, A., Harris, R., Hjort, N., y Jones, M. C., (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, **85**, 549–559, 1998.

Ghosh, A. y Basu, A. (2016). Robust estimation in generalized linear models: the density power divergence approach. Test, **25**(2), 269-290.

Ghosh, A., Jaenada, M. y Pardo, L. (2021) Robust adaptive variable selection in ultra-high dimensional linear regression models. arXiv:2004.05470

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58, 267–288.

Zou, H. y Hastie, T. (2006). The adaptive lasso and its oracle properties. Journal of American Statistical Association **101**, 1418–1429.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ASIGNACIÓN DE DATOS ECONÓMICOS AO DIRECTORIO DE EMPRESAS E UNIDADES LOCAIS PARA A OBTENCIÓN DO PRODUTO INTERIOR BRUTO MUNICIPAL

Teijeiro Campo, T.[1], Calvo Ocampo, E.,[1]Jácome Rodríguez, R.[1], Suárez Morao, M.[1], Vilar Cruz, C.[1]

[1] Instituto Galego de Estatística

**RESUMO**

O obxectivo deste traballo é a obtención dun directorio de empresas e establecementos con información do valor engadido xerado por cada unidade, que permita obter estimacións do valor engadido por rama de actividade xerado en cada concello. Estas estimacións contribúen á mellora da operación estatística "Produto interior bruto municipal" e achegan robustez ás estimacións do Sistema de Contas Económicas do IGE.

**Palabras e frases chave:** Produto interior bruto, establecemento, rama de actividade, matching.

## 1. INTRODUCIÓN

A demanda de información económica cun maior nivel de desagregación territorial e, en concreto, por concello, ten medrado de xeito significativo nos últimos anos. A ausencia de información económica agregada, comparable e coherente co Sistema de Contas Económicas de Galicia era unha das eivas da produción estatística do Instituto Galego de Estatística.

O obxectivo da operación estatística "Produto interior bruto municipal" é a obtención de estimacións municipais e comarcais do Produto interior bruto (PIB) ademais de achegar información da estrutura produtiva das comarcas e dos concellos de maior tamaño. Para acadar estes fins utilízase un amplo e variado conxunto de información estatística de base que garda coherencia coa información macroeconómica que o Instituto Galego de Estatística (IGE) difunde para a Comunidade autónoma.

A nova base contable ("Revisión Estatística 2019") presentaba unha oportunidade para realizar melloras nos procedementos internos de estimación e aplicar os cambios metodolóxicos introducidos na publicación das Contas económicas anuais (IGE, 2019b) e no Marco Input-Output (IGE, 2019a).

A economía de cada concello está formada por todas as unidades institucionais que teñen un centro de interese económico predominante no territorio económico do concello, que asimilaremos ao territorio xeográfico (IGE, 2020).

O PIB é un indicador da actividade total nun territorio económico. Hai tres formas de medir o PIB a prezos de mercado, porén nas estimacións municipais prima o enfoque da produción, ou oferta. Segundo este enfoque o PIB será a suma de todos os valores engadidos (VEB) de todas as actividades de produción de bens e servizos, máis os impostos sobre os produtos menos as subvencións sobre os produtos.

O punto de partida para as estimacións do PIB municipal é a información dispoñible nas Contas económicas anuais, nas que se estiman as contas de produción e explotación para 72 ramas de actividade. Para cada unha delas desagregouse o valor engadido bruto por concello utilizando diversos métodos de estimación en función das distintas fontes estatísticas dispoñibles.

Porén, a salvagarda do segredo estatístico é a principal causa de que non sexa viable a difusión do PIB municipal coa mesma desagregación que o dato autonómico. A calidade das fontes estatísticas de base tamén xogan un papel relevante nestas limitacións. Así, só para os concellos de máis de 10.000 habitantes se publica unha difusión do PIB a catro grandes sectores : primario (que inclúe a agricultura, gandaría, silvicultura, pesca e acuicultura), enerxía e industria, construción e servizos. As comarcas permiten unha desagregación máis ampla e publícase información de 12 ramas de actividade, excepto en determinadas agregacións de actividade que, para garantir o segredo estatístico, censúranse determinados datos. En todo caso, a desagregación mínima ofrecida no caso das comarcas é de 9 agrupacións de actividade.

| **Ramas de actividade** | **Códigos CNAE-09 que se inclúen:** | **Difusión provincial e comarcal** | **Difusión concellos de máis de 10.000 habitantes** |
|---|---|---|---|
| **PRIMARIO** | **SECCIÓN A Ramas: 01-03** | **X** | **X** |
| **ENERXÍA E INDUSTRIA** | **SECCIÓNS: B,C,D,E** | **X** | **X** |
| Industria agroalimentaria | 10-11 | **X** | |
| Madeira, papel e mobles | 16-17;31 | **X** | |
| Industria extractiva e pesada | 05-09, 13-15,18-33 | **X** | |
| *Industrias extractivas e fabricación de outros produtos minerais non metálicos* | *05-09,23* | ***X*** | |
| *Metalurxia e produtos metálicos, electrónicos, eléctricos e maquinaria* | *24-28* | ***X*** | |
| *Fabricación de material de transporte e grandes reparacións industriais* | *29-30; 33* | ***X*** | |
| *Resto da industria* | *13-15; 18-22; 32* | ***X*** | |
| Enerxía, subministro de auga e xestión de residuos | 35-39 | **X** | |
| **CONSTRUCIÓN** | **SECCIÓN F Ramas 41-43** | **X** | **X** |
| **SERVIZOS** | **SECCIÓNS: G- T** | **X** | **X** |
| Comercio, transporte e hostalería | 45-56 | **X** | |
| Actividades de información, financeiros, inmobiliarios e profesionais | 58-82 | **X** | |
| Administracións Públicas, Educación, Sanidade e outros servizos | 84-97 | **X** | |

Táboa 1. Difusión sectorial da operación estatística 3902-03-OE05 Produto interior bruto municipal (IGE, 2020)

O Sistema Europeo de Contas, SEC-2010, (Eurostat, 2013) introduce no capítulo sobre as contas rexionais os métodos de rexionalización das contas a escala nacional. Na estimación do VEB municipal, utilizáronse estes métodos, tendo en conta que se parte dos agregados para a economía de Galicia publicados nas Contas Económicas Anuais e se busca obter determinados agregados para espazos económicos máis reducidos (concellos e comarcas). A adaptación á estimación dos agregados municipais dos métodos de rexionalización que propón o SEC (SEC-2010; 13.24) sería a seguinte:

a) Métodos ascendentes. Para unha determinada rama de actividade, este método supón utilizar a información das unidades residentes no concello e ir agregando a mesma ata estimar o VEB xerado no concello.

b) Métodos descendentes. Este método implica distribuír por concello a cifra do VEB estimado para Galicia nunha rama de actividade utilizando un indicador de reparto por concello que reflicta o máis exactamente posible o VEB xerado.

c) Métodos mixtos. O método descendente non se presenta, polo xeral, en estado puro, polo que debe considerarse tamén a posibilidade de utilizar métodos mixtos, nos que se combinan métodos ascendentes cos descendentes.

A información dispoñible para cada rama de actividade determinará en última instancia a utilización dun ou doutro método, sempre dando prioridade ao método ascendente, que denominaremos "novo procedemento xeral" (NPX).

O novo procedemento xeral para obter estimacións por municipio nas diferentes ramas de actividade busca construír o valor engadido total en cada municipio como suma das unidades locais (establecementos produtivos) presentes nese termo municipal (método ascendente). E para iso combina a información das empresas dispoñible no IGE e que resulta pertinente para os fins desta operación (información económica que permita a súa clasificación xeográfica e por actividade).

Dicimos busca construír, porque non existe información individual por establecemento para obter estimacións municipais. Por este motivo, o valor engadido bruto por rama de actividade para o conxunto de Galicia é a referencia en cada rama e búscase a fórmula de asignar valor engadido a cada concello mantendo a coherencia do total da comunidade autónoma. Veremos nesta comunicación como esta idea vai cambiando progresivamente.

## 2. ANTECEDENTES

Nos seus inicios o procedemento de estimación do PIB municipal centrouse na busca de indicadores para distribuír por concellos o valor engadido bruto de cada rama de actividade publicado nas Contas económicas anuais; polo tanto, encadrábase dentro dos métodos de estimación descendentes.

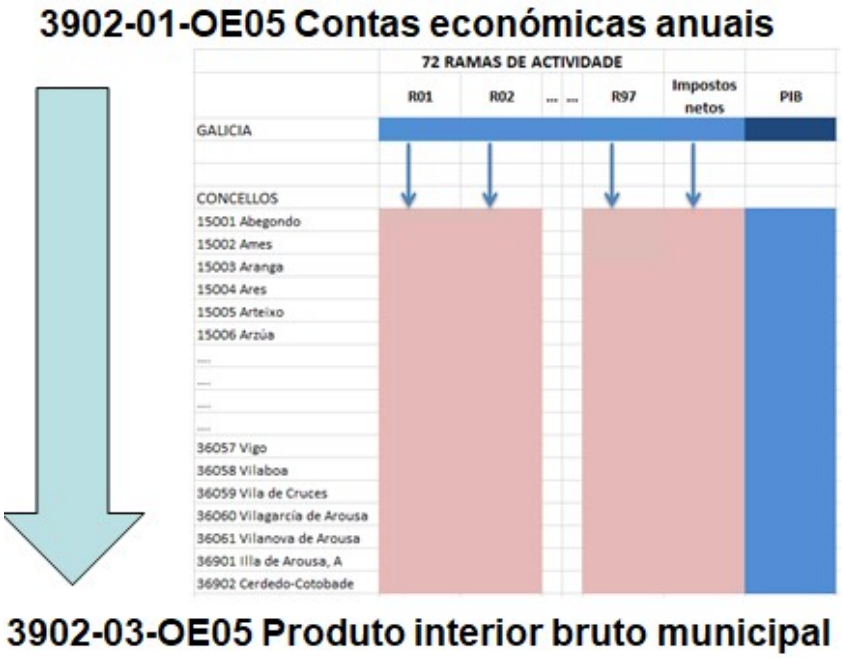



Figura 1. Esquema anterior procedemento de estimación no PIB municipal

Dentro deste esquema unha boa parte das estruturas por concello das ramas produtivas procedían dun procedemento de estimación que se denominará "anterior procedemento xeral" de estimación (APX) e que consistía en partir dos elementos mostrais das enquisas estruturais de empresas (sector industrial e de servizos) e combinar esta información con estimacións do VEB por produtividades do resto de unidades. O primeiro paso era calcular a suma do emprego por rama e estrato de emprego do establecemento (obtido do aproveitamento da información do Directorio Central de Empresas do INE intercambiado co IGE para a mellora do noso Directorio de Empresas e Unidades Locais) das empresas non seleccionadas nas mostras das enquisas estruturais. Para imputar un valor engadido a ese emprego total multiplicábase pola produtividade da rama e estrato calculado para o total de Galicia coas unidades mostrais das enquisas estruturais. Baixo esta filosofía, o obxectivo era unha estimación da estrutura territorial de cada rama de actividade, pero a coherencia cos totais estimados para o total de Galicia nas *Contas económicas anuais* era un obxectivo secundario. Tampouco era un obxectivo prioritario dispor de información económica individual das empresas non seleccionadas nas mostras, senón que se estimaban agregados de empresas por rama e estrato.

Nas ramas de actividade non cubertas polas estatísticas estruturais de empresas non se utilizou, nin se utiliza polo momento, o "procedemento xeral" (PX), xa sexa pola ausencia de información individual ou pola natureza particular da estimación da rama nas Contas anuais.

No seguinte gráfico represéntase a distribución do VEB segundo o procedemento de estimación utilizado.

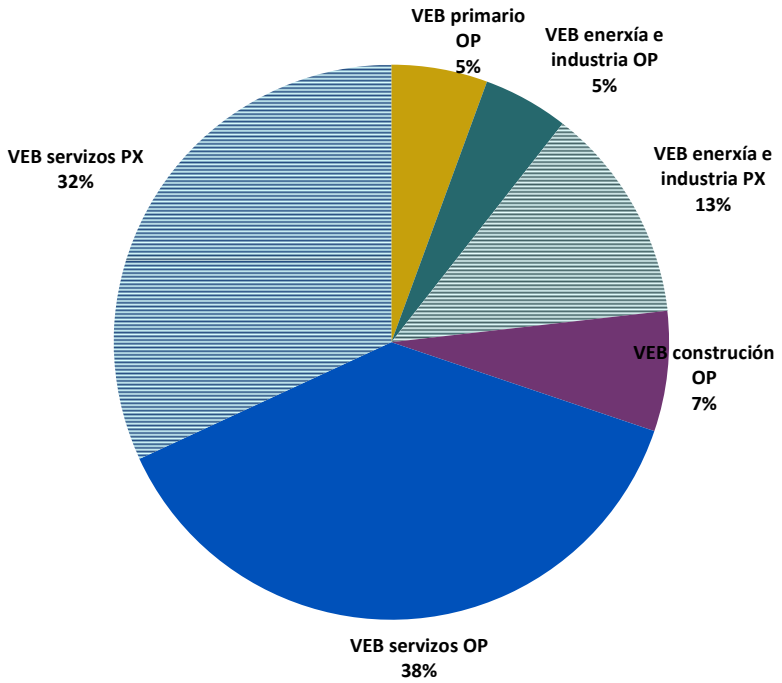



Gráfico 1. Distribución do valor engadido bruto xerado en Galicia por sector de actividade e tipo procedemento de estimación. Ano 2016.

Como se pode comprobar no Gráfico 1, o procedemento xeral non se emprega para estimar todas a ramas de actividade; só se utiliza para estimar unha parte dos servizos e unha parte da enerxía e industria. Así, no ano 2016 a porcentaxe do VEB de Galicia estimada a través do PX foi o 45% (un 13% corresponde a unha parte da enerxía e industria e un 32% aos servizos). O 55% do VEB restante estímase a partir doutros procedementos (OP), xa que non utiliza a información das enquisas estruturais a empresas como punto de partida das estimacións.

## 3. PROCEDEMENTO XERAL

O novo procedemento xeral para obter estimacións por municipio nas diferentes ramas de actividade busca construír o valor engadido en cada municipio como agregación das unidades locais (establecementos produtivos) presentes nese termo municipal (método ascendente). Polo tanto, o principal obxectivo deste traballo é conseguir un listado de establecementos de Galicia con información da súa rama de actividade económica, concello, emprego e valor engadido bruto.

O cálculo do valor engadido realízase a partir de datos contables das empresas recollidos nas fontes que se relatan nesta sección. Estímase por separado a "produción" e os "consumos intermedios" e o valor engadido é a diferenza entre estes conceptos. A produción recolle fundamentalmente a facturación da empresa debida á súa actividade principal ( a través da variable contable "Importe neto da cifra de negocios") pero tamén outros ingresos derivados de actividades accesorias ("Ingresos accesorios e outros de xestión corrente") ou ingresos imputados como aqueles derivados de actividades que redundan nun maior valor do activo da empresa ("Traballos realizados pola empresa para o seu activo"). Os consumos intermedios recollen os bens e servizos utilizados no proceso produtivo que, contablemente, se materializa en "Aprovisionamentos" e "Servizos exteriores" fundamentalmente.

O punto de partida para a realización deste proxecto foi o Directorio de establecementos de Galicia do IGE, que proporciona datos de localización, número de asalariados e actividade principal dos establecementos.

Non obstante, cómpre resaltar que o proceso de construción da táboa fíxose de forma secuencial, é dicir, a información económica incorporouse seguindo unha xerarquía nas fontes e previo cruce coa información dispoñible no directorio do IGE para contrastar a información de localización e nalgún caso da actividade dos establecementos produtivos.

O procedemento consiste esencialmente na integración neste directorio dos datos económicos das fontes de información dispoñibles cunha serie de limitacións derivadas da natureza destas últimas.

As fontes que se consideraron foron as seguintes:

Estadística Estructural a Empresas (EEE). A información recollida nesta estatística, elaborada polo *Instituto Nacional de Estadística*, permite estimar distintos agregados macroeconómicos (como a produción, os consumos intermedios, emprego, ...) en termos do SEC-2010, a partir dos datos contables das empresas (INE, 2021).

No caso do sector industrial, proporciona datos de localización por concello tanto da sede social da empresa como dos seus establecementos e; no caso do sector servizos, achega datos agregados por comunidade autónoma, é dicir, aquelas empresas con máis dun establecemento en Galicia estarán agregadas nun único rexistro e será necesario combinar esta fonte co Directorio de empresas e unidades locais do IGE para asignar os datos de valor engadido bruto e emprego a cada establecemento.

A mostra da EEE para Galicia recolle datos duns 4.700 establecementos industriais e de 9.000 empresas do sector servizos.

Estadística Minera. Esta estatística, elaborada polo *Ministerio para la Transición Ecológica y el Reto Demográfico*, proporciona información para cada explotación mineira da súa localización, do valor da produción e dos custos de produción, o que permite o cálculo do valor engadido bruto xerado por explotación. Recolle información dunhas 200 explotacións galegas.

Encuesta de Estructura de la Industria de la Construcción (EEC). Esta enquisa anual é responsabilidade do *Ministerio de Transportes, Movilidad y Agencia Urbana* e está dirixida a empresas (sociedades e autónomos) cuxa actividade principal é a construción. Ten como obxectivo o coñecemento das principais macromagnitudes do sector da construción. Achega información contable por empresa e, para obter datos por comunidade autónoma, dispón de catro variables: porcentaxe de vendas, emprego medio, custo laboral e número de locais por comunidade autónoma. Os datos permiten calcular o valor engadido bruto e o emprego para Galicia de cada empresa recollida na mostra, entre 1.000 e 1.100 unidades cada ano.

Base de datos de información depositada polas empresas nos rexistros mercantís de Galicia (SABI). Esta fonte proporciona para cada empresa, información da conta de perdas e ganancias, o que permite calcular o valor engadido bruto da empresa nos termos explicados anteriormente. Ten datos de aproximadamente 55.000 sociedades con sede social en Galicia, e máis de 800.000 sociedades de toda España.

A principal dificultade á hora de explotar esta fonte radica en discriminar se a actividade da empresa con sede social en Galicia se pode imputar en exclusiva á nosa Comunidade (e aos seus concellos) ou se, pola contra, a empresa tamén exerce actividade fóra de Galicia.

Por outra parte, tamén será necesario combinar esta fonte co Directorio de empresas e unidades locais do IGE para asignar os datos de valor engadido bruto e emprego a cada establecemento.

No proceso de asignación do valor engadido bruto a cada establecemento pódense diferenciar dúas fases:

- Fase 1: incorporación de rexistros correspondentes ás fontes anteriormente citadas (EEE, MINERA, ECC, SABI), de forma xerárquica e mediante cruces co Directorio de establecementos do IGE para confirmar a localización dos establecementos produtivos e nalgún caso incluso para incorporar os datos dos establecementos.

- Fase 2: incorporación dos datos dos establecementos do Directorio de establecementos do IGE non incluídos a través dalgunha das fontes anteriores e imputación do VEB de cada establecemento.

A variable emprego do establecemento é fundamental en todo o proceso de asignación do VEB por establecemento, tanto nalgúns casos, pola necesidade de utilizar esta variable para distribuír o VEB da empresas entre os seus establecementos, como noutros que se precisa para imputar o VEB do establecemento coas produtividades de referencia.

Os datos de emprego dos establecementos do Directorio do IGE obtéñense a partir dos datos de afiliados á Seguridade Social e inclúen polo tanto só o emprego asalariado. Co fin de evitar que existan establecementos con emprego igual a 0, o que dificultaría a asignación do VEB, realizouse un pequeno axuste sobre o emprego dos establecementos do Directorio do IGE:

- No caso de sociedades con dato de emprego igual a 0 no Directorio do IGE, imputóuselle o valor 1 ao emprego asalariado. Considerouse neste casos que a empresa terá polo menos un empregado e, seguindo a metodoloxía SEC-2010, que ao tratarse dunha sociedade será emprego asalariado.

- No caso de persoas físicas, imputóuselle o valor 1 ao emprego non asalariado. Asumiuse neste caso que o autónomo é un empregado da empresa e ao tratarse dunha empresa non constituída en sociedade será emprego non asalariado.

Fase 1

A **primeira fonte considerada** é a EEE porque permite calcular o valor engadido bruto por establecemento (no caso da industria) e o valor engadido bruto da empresa en Galicia (no caso dos servizos) e ademais é a fonte básica de información na elaboración das Contas económicas de Galicia.

O cruce dos datos desta fonte cos do Directorio de establecementos do IGE permite:

- Asignar o VEB por establecemento nos caso nos que na EEE só figura o dato para a empresa.
- Mellorar o dato da localización do establecemento que aparece na EEE co Directorio do IGE.
- Contrastar os datos da actividade económica do establecemento recollida en ambas as dúas fontes.
- Incorporar empresas e establecementos non recollidos no Directorio do IGE.

No proceso de matching dos datos da EEE co Directorio de establecementos do IGE apareceron unha serie de obstáculos que foi necesario solventar, como a existencia de empresas na EEE que non figuraban no Directorio do IGE ou as diferenzas existentes entre ambas as dúas fontes tanto no número de establecementos, como na localización e a actividade destes últimos.

A información final obtida a través deste matching é o resultado dun proceso de conciliación de fontes.

No número de establecementos seguiuse un criterio de máximos, incluíndo o número máximo de rexistros para cada empresa. Así, nos casos nos que a EEE proporcionaba maior detalle dos establecementos que o Directorio, incorporouse o número de rexistros da EEE e, nos outros casos os do Directorio.

No caso da actividade económica do establecemento respectouse a actividade da EEE nos casos nos que nesta última fonte se dispoñía deste dato, xa que aínda que sempre figura a actividade económica da empresa, existen numerosos casos (fundamentalmente no sector servizos) para os que non se dispón da información desagregada por establecemento. Nesta última situación asignouse a actividade económica dos establecementos do Directorio do IGE.

Polo que respecta á localización dos establecementos, respectouse en todo caso o concello que figura no Directorio, excepto naqueles establecementos que se incorporaron no listado final procedentes da EEE e nos que non foi posible unir con datos de establecementos do Directorio do IGE; ben por rexistrar máis establecementos a EEE que os recollidos no Directorio ben por non poder establecer relacións biunívocas entre establecementos de ambas as dúas fontes ao non existir un identificador único de establecemento.

O emprego dos establecementos que se incluíu no listado final foi na gran maioría dos rexistros o asociado a cada establecemento do Directorio do IGE logo de realizar os axustes mencionados anteriormente. Ao igual que no caso da localización, existen algúns casos nos que a cifra do emprego é a que figura na EEE, ao incluír directamente no listado final rexistros da EEE que non se puideron vincular con establecementos do Directorio.

A EEE proporciona o VEB da empresa e nalgúns casos dos establecementos. Ao longo do proceso de matching co DIRIGE houbo que asignar o VEB da empresa entre os distintos establecementos, fundamentalmente nos casos nos que a información do establecemento non estaba dispoñible na enquisa. O emprego do establecemento foi a variable utilizada para distribuír o VEB da empresa entre os distintos establecementos.

A **segunda fonte considerada**, a Estadística Minera, ao igual que a EEE tamén proporciona información do valor engadido bruto por establecemento, pero só cubre a actividade extractiva.

Incorporouse ao listado resultante da EEE os rexistros correspondentes á Estadística Minera de empresas que non se incluíran xa a través da EEE.

**A EEC é terceira fonte** que se incorporou no proceso de integración de fontes. Nesta fase engadiuse información relativa a aquelas empresas que non estaban incluídas na táboa final por algunha das fontes anteriores (EEE ou Estadística Minera).

A partir dos datos da EEC calculouse o valor da produtividade da empresa e determinouse o valor engadido bruto dos establecementos que figuran para esa empresa no Directorio de establecementos do IGE a partir desta produtividade e do valor do emprego asociado ao establecemento no Directorio.

A **última fonte** que se considerou para incorporar á táboa final de establecementos foi a Base de datos de información depositada polas empresas nos rexistros mercantís de Galicia (SABI).

Incluíronse no listado final só empresas que non se incluíran xa por algunha das outras fontes (EEE, Estadística Minera, EEC).

Esta fonte só permite calcular o valor engadido bruto da empresa, polo que cómpre ter en conta se a empresa exerce toda a súa actividade en Galicia ou se desenvolve a súa actividade en Galicia e no resto de España.

Para determinar o ámbito da actividade realizouse unha comparativa entre o emprego rexistrado para a empresa na base de datos de SABI e o rexistrado no Directorio do IGE e se este último era maior ou igual que o de SABI, asumiuse que se trataba de empresas que desenvolvían toda a súa actividade en Galicia.

Para as empresas con actividade en Galicia e fóra de Galicia, consideráronse todos os establecementos da empresa do Directorio do IGE e o VEB asociado a cada establecemento calculouse a partir da produtividade da empresa en SABI e do emprego do establecemento no Directorio.

Para as empresas que desenvolven toda a súa actividade en Galicia, o valor engadido bruto da empresa en SABI repártese entre os establecementos utilizando a distribución do emprego do Directorio do IGE.

Fase 2

Cos datos de todos os establecementos incorporados na fase 1 calcúlase a produtividade por rama de actividade e estrato. Este valor da produtividade, xunto co dato do emprego do establecemento do Directorio de establecementos do IGE, permite imputar a cada un dos establecementos non incluídos na Fase 1 un valor engadido bruto obtido como:

$$VEB_{establecemento} = Produtividade_{Rama,Estrato} * Emprego_{Dirige}$$

Como resultado de todo este proceso de integración de fontes obtívose finalmente unha táboa de datos na que cada rexistro se corresponde cun establecemento e se recollen ademais as seguintes variables asociadas a este último: concello, actividade económica, emprego e valor engadido bruto.

No proceso de construción da táboa incorporouse tamén información sobre a fonte de cada dato, xa que como se mencionou anteriormente, os rexistros proceden na maior parte dos casos dun proceso de selección de fontes.

Este novo procedemento incorpora, a diferenza do anterior, novas fontes de información como a Estadística Minera, a EEC e a procedente de SABI, o que permite:

- Incrementar o número de rexistros que se estiman directamente e mellorar polo tanto o cálculo das produtividades por rama e estrato.
- Reducir o número de rexistros estimados a partir da produtividade por rama de actividade e estrato.

Por outra parte, o feito de dispoñer de rexistros individuais posibilita, no momento no que se poda identificar de forma única a cada establecemento e ademais se dispoña dunha serie temporal, realizar outros tipos de estimacións que permitan incorporar a información pasada de cada establecemento para estimar o valor engadido bruto no último ano considerado.

## 4. RESULTADOS

O novo procedemento permite testar as estimacións que para Galicia se realizan nas Contas económicas anuais. Este obxectivo, secundario antes de implantar este proceso, convértese en prioritario. Xa non se entenden de xeito independente as dúas operacións estatísticas, senón que deben desenvolverse en paralelo e as estimacións internas do PIB municipal por rama de actividade deben servir para mellorar e contrastar as estimacións para Galicia antes da súa difusión. É dicir, estase a desenvolver o método ascendente, que permitirá obter estimacións para Galicia como suma de cada un dos establecementos produtivos de cada rama de actividade, que ademais se terán localizados nos municipios correspondentes.

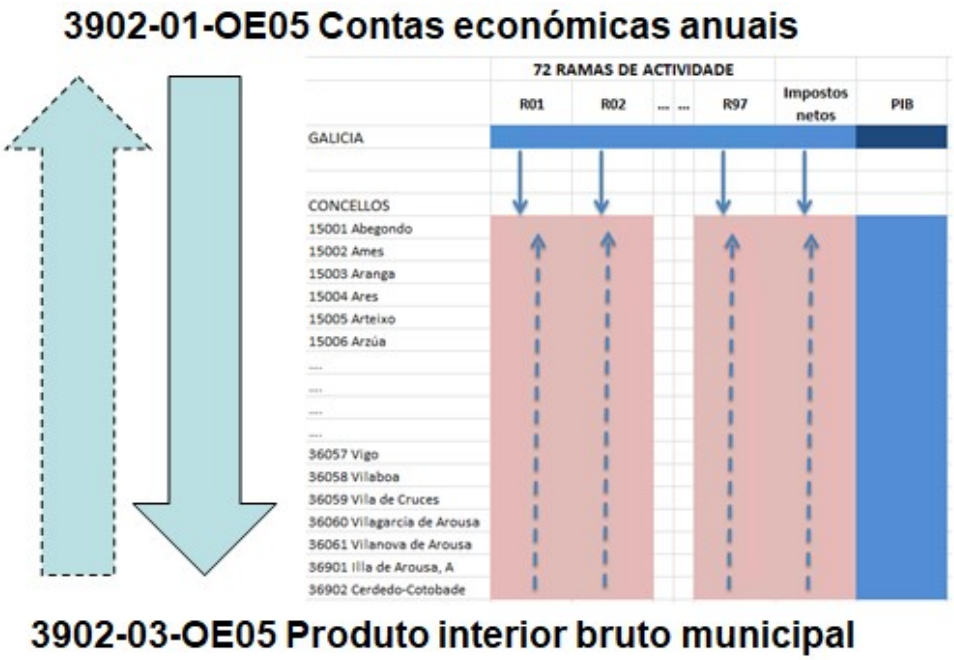



Figura 2. Esquema novo procedemento de estimación no PIB municipal

Como vimos no segundo bloque, isto será posible naquelas ramas que se estiman mediante o procedemento xeral.

Na seguinte gráfica expóñense as estimacións da EEE xunto coa estimación final do NPX para industria e para servizos[1].

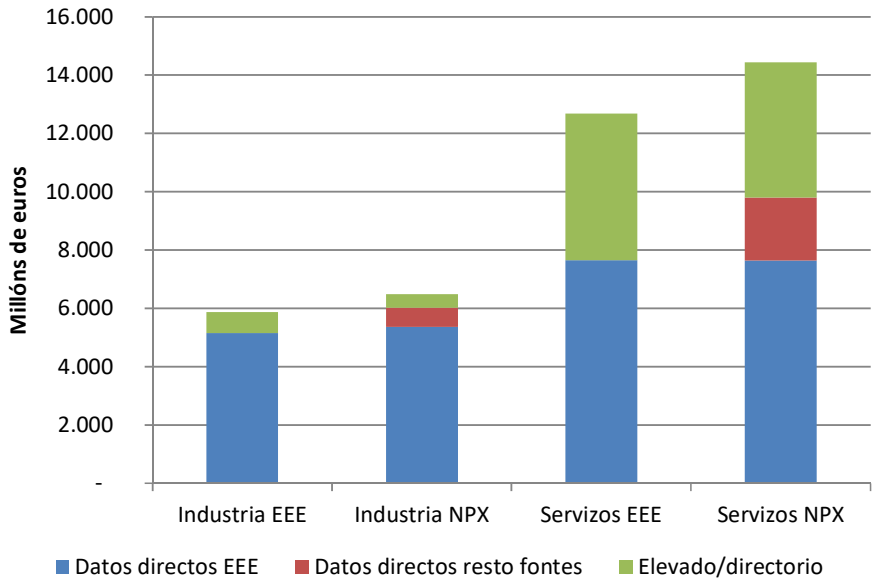



Gráfico 2: Valor engadido bruto nas ramas de actividade distribuídas por concellos no novo procedemento xeral. Ano 2016.

[1] Recordemos que o valor engadido "municipalizado" co procedemento xeral representa o 45% do valor engadido total.

Representamos como "Industria EEE" e "Servizos EEE" o valor engadido que procede da explotación da "*Estadística Estructural de Empresas".* A cor azul representa a suma dos datos mostrais sen elevar e a cor verde a parte resultante da elevación. Ao ser unha mostra, cada elemento ten un factor de elevación asociado, e a cor verde representa unha estimación do valor engadido das unidades non seleccionadas na mostra.

Representamos como "Industria NPX" e "Servizos NPX" o valor engadido estimado tras aplicar o procedemento de asignación de VEB aos establecementos produtivos nestas ramas de actividade. Unha parte deste valor engadido provén das novas fontes (cor vermella) que nos permiten, mediante matching, asignar un valor engadido e a parte en cor verde representa o total asignado tras as imputacións, rexistro a rexistro, dunha produtividade media ao emprego do directorio.

Debemos resaltar os seguintes resultados:

- No novo procedemento xeral de estimación ao dispor de información dun maior número de unidades mostrais (procedentes doutras fontes de información diferentes das estatísticas estruturais) obtéñense estimacións máis robustas. Esta mellora apreciase sobre todo nos servizos onde os factores de elevación nas estatísticas estruturais son máis elevados. A barra verde en "Servizos EEE" é moi superior á "Industria EEE". O valor engadido industrial está concentrado en menos unidades produtivas (establecementos) que o valor engadido no sector servizos.
- O maior número de unidades mostrais procede das novas fontes (Minera, EEC e SABI) e é especialmente relevante no sector servizos nos que o NPX achega máis información individual que somos capaces de clasificar por rama e concello e de asignar dato de VEB e emprego.
- Os niveis estimados finalmente son máis elevados, xa que o Directorio de Empresas e Unidades Locais do IGE é máis exhaustivo que o directorio de partida da EEE. Nótese que isto ocorre sobre todo no sector servizos.

## 5. CONCLUSIÓNS

Concluímos cunha matriz DAFO (debilidade, ameazas, fortalezas e oportunidades) do procedemento que describimos nas seccións anteriores.

| Análise Interno | Análise externo |
|---|---|
| **D**EBILIDADES<br>- O novo procedemento xeral de estimación (NPX) non é extensible a todas as ramas de actividade da economía galega, ao ser tamén diferente a información dispoñible para cada unha delas.<br>- Cambios na forma xurídica das empresas, fusións, grupos empresariais, ... | **A**MEAZAS<br>- Dependencia do Directorio de establecementos do IGE, que proporciona datos de localización, número de asalariados e actividade principal dos establecementos.<br>- A normativa actual non obrigue á existencia dun código único de establecemento. |
| **F**ORTALEZAS<br>- Contraste da calidade das contas anuais. Nas contas económicas estimamos macromagnitudes por rama de actividade, no PIB municipal debemos estimalas por concello e rama.<br>- Mellora das estimacións ao incrementar o número de unidades con información directa e reducir o número de unidades estimadas por produtividade rama-estrato.<br>- A disposición de información de carácter lonxitudinal por establecemento permite introducir novas formas de estimación. | **O**PORTUNIDADES<br>- Os resultados por establecemento empresarial permiten unha rápida valoración económica de grupos empresariais (asociacións, clusters, ...).<br>- Unión con outras fontes de información por establecemento, non só de carácter económico. |

## REFERENCIAS

EUROSTAT (2013): *Sistema Europeo de Cuentas Nacionales y Regionales de la Unión Europea. SEC 2010*. Reglamento UE Nº549/2013 del Parlamento Europeo y del Consejo

IGE (2019a). Nota sobre a "revisión estatística 2019" nas operacións do Sistema de Contas Económicas de Galicia http://www.ige.eu/estatico/pdfs/s3/metodoloxias/Nota_Revision_Estatistica_2019_gl.pdf

IGE (2019b). Contas económicas anuais. Revisión estatística 2019.
http://www.ige.eu/estatico/pdfs/s3/metodoloxias/met_contas_economicas_gl.pdf

IGE (2020). Produto Interior Bruto municipal Revisión estatística 2019. Metodoloxía.
http://www.ige.eu/estatico/pdfs/s3/metodoloxias/met_PIB_municipal_RE19_gl.pdf

INE (2021). Estadística Estructural de Empresas. Metodología.
https://ine.es/metodologia/t37/metodologia_eee2019.pdf

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# INDICADORES SUMARIOS DALGÚNS TABULADORES OU DAS SÚAS COMPOÑENTES

Carlos L. Iglesias Patiño[1]

[1] Instituto Galego de Estatística

**RESUMO**

Introducimos indicadores para sintetizarmos funcións, de conxunto ou argumento nominal, que serven para tabular. Interpretan matematicamente índices ou conceptos socioeconómicos o que permite empregalos noutros contextos e mesmo favorecen a visualización e tratamento da información na estatística pública.

**Palabras e frases chave:** tabuladora, indicador de Herfindahl-Hirschman, primacía, número efectivo de clases

## 1. INTRODUCIÓN

O tratamento conxunto da tabela estatística e non só de cada unha das súas celas pode ter moito interese tanto teórico como práctico. No fondo é un proceso semellante á xeneralización da ecuación á función pasando polo polinomio. Por iso introducimos os conceitos de descritor aplicativo (Glez. Manteiga e Iglesias, 2018) ou de tabulador (Iglesias, 2019) e aplicámolos á macro-depuración (Iglesias e López Vizcaíno, 2018).

A visión da liña dunha tabela como unha función de conxunto ou argumento nominal vai permitir empregarmos útiles adoito desenvolvidos noutras ramas das matemáticas ou doutras disciplinas matematizadas. O resultado conducirá a uns indicadores que han resumir esa liña e xeneralizan algúns dos xa empregados na administración pública (IGE e DRN-INE, 2000 e mais U.S. Department of Justice, 2019) ou mesmo nas ciencias sociais ou nas da vida.

Á par deste proceso de abstracción, procuraremos concretar con exemplos tanto aritméticos como gráficos que amosen a utilidade desta perspectiva auxiliándonos do R na súa construción.

## 2. NORMAS DAS TABULADORAS

Entendemos por tabuladora unha aplicación que serve para tabular inclúe polo tanto os tabuladores e as súas compoñentes (Iglesias, 2019). Habemos definir a norma-r dela a partirmos da seguinte igualdade

$$\|T(\cdot)\|_r^r = \sum_{c\in\mathcal{C}} |T(B_c)|^r$$

sendo $\mathcal{C}$ un conxunto de literais (v.g. unha escala nominal) tal que $\#\mathcal{C} = C$ e $\bigsqcup_{\mathcal{C}} B_c$ unha partición do colectivo que estamos a estudar. Nomeadamente, habemos tratar os casos $r = 1$ e $2$ e mais a norma do supremo

$$\|T(\cdot)\|_\infty = \sup_{c\in\mathcal{C}} |T(B_c)|.$$

Se a aplicación $T(\cdot)$ é un subtotalizador $Y_+(\cdot)$ e a variábel $Y$ é non-negativa, $\|Y_+(\cdot)\|_1$ coincide co total dela no colectivo. Nesta situación habitual en estatística pública, podemos construír unha nova aplicación tabuladora $\|Y_+(\cdot)\|_1^{-1} Y_+(\cdot)$ que denominaremos partilla cuxa norma-1 é a unidade. Cada imaxe é a cota que lle corresponde a esa clase, representante ou literal, $q(B_c)$. Un caso particular interesante é cando consideramos o recontador, $N(\cdot)$, entón $\|N(\cdot)\|_1^{-1} N(\cdot)$ é a aplicación de frecuencias (AF). Outro caso particular é $\|R_1(\cdot)\|_1$ cando $Y = X = 1$ (Iglesias, 2019) que constitúe un primeiro indicador do variado ou mudábel que é o colectivo, de aí que o denominemos indicador de diversificación ($ID$), isto é (i.e.), o número de clases que non están baleiras ou non son nulas.

Se $Y$ é xeral convén introducirmos as partes positiva e negativa de $Y$, $Y^+ := \max\{Y, 0\}$ e $Y^- := \max\{-Y, 0\}$ donde $Y = Y^+ - Y^-$ e $|Y| = Y^+ + Y^-$. A positiva $Y^+$ resulta medio aritmético de $Y$ e $|Y|$. A partirmos delas podemos introducir dúas tabuladoras $Y_+^+(B_c)$ e $Y_+^-(B_c)$ que acumulan os valores positivos e negativos de $Y$ respectivamente. Polo tanto podemos ver o subtotalizador como saldo $Y_+(\cdot) = (Y_+^+ - Y_+^-)(\cdot)$ e considerar a tabuladora $|Y|_+(\cdot) := (Y_+^+ + Y_+^-)(\cdot)$ e mais o saldo relativo $[(|Y|_+)^{-1}(Y_+^+ - Y_+^-)](\cdot)$. O saldo global é $\|Y_+^+(\cdot)\|_1 - \|Y_+^-(\cdot)\|_1$. A seguir, agás comentario en contra, $Y$ representará unha variábel non-negativa.

O cadrado da norma-2 dun subtotalizador é a suma de cadrados (SC) dos seus subtotais compoñentes cando $C$ é finito como sempre acontece en estatística pública. No caso da partilla, estamos a calcular $\|q(\cdot)\|_2^2$ nas aplicacións dun subconxunto da esfera unidade coa norma-1, $\|q(\cdot)\|_1 = 1$. A SC da partilla ha ser menor ca súa norma-1, porque $q^2(B_c) < q(B_c)$ agás no caso de que $q^2(B_c) = q(B_c) = 1$, $ID = 1$, en que ambas as dúas coinciden. Isto é, a norma-2 seica está relacionada cunha medida de concentración nesta situación. Por outra banda, esta norma pode desempeñar un papel complementario no tratamento de variábeis xerais, i.e., que poidan tomar valores negativos.

Cando consideramos $\|Y_+(\cdot)\|_\infty = \sup_c Y_+(B_c)$ podemos definir outro indicador sumario como a clase (representante ou literal) onde se acada ese supremo –máximo en estatística pública–. A esta clase habémola denominar primacial e a $\|Y_+(\cdot)\|_\infty / \|Y_+(\cdot)\|_1$ primacía dela. Tamén pode esta terceira norma ter máis senso cos tabuladores, $R_g$, agás no devandito caso do $ID$ para $g = 1$.

## 3. INDICADORES

Ademais do total, do $ID$ e da primacía, habemos construír más indicadores e operadores valéndonos desas normas. Por exemplo, podemos construír o seguinte indicador de concentración

$$HH = \mathrm{HH}[Y_+(\cdot)] = \left(\frac{\|Y_+(\cdot)\|_2}{\|Y_+(\cdot)\|_1}\right)^2 = \left\|\frac{Y_+(\cdot)}{\|Y_+(\cdot)\|_1}\right\|_2^2 = \|q(\cdot)\|_2^2$$

A notación provén de que semella o índice de Herfindahl e Hirschman. Este indicador verifica que $C^{-1} \le \mathrm{HH}[Y_+(\cdot)] \le 1$. A primeira igualdade acádase na alícuota, cando a variábel $Y$ está equi-repartida, e a segunda, xa foi tratada, corresponde a concentración máxima. Se o número de clases $C$ é grande a minorante aproxímase a 0.

Pode interpretarse *HH* en termos da variancia relativa ou coeficiente de variación das cotas, abusándomos da notación,

$$HH = \frac{1 + C_V^2[q(\cdot)]}{C}$$

porque $\mathrm{D}^2[q(\cdot)] = C^{-1}\sum_c q^2(B_c) - \left(C^{-1}\sum_c q\,(B_c)\right)^2 = C^{-1}HH - C^{-2}$. O coeficiente de variación das cotas pode interpretarse, a súa vez, como a taxa de variación do *HH* con base na alícuota, que era o menor valor que podía tomar

$$C_V^2[q(\cdot)] = \frac{[C^{-1}HH - C^{-2}]}{C^{-2}} = \frac{HH}{C^{-1}} - 1 = C\cdot HH - 1$$

O complemento a unidade del $1 - HH = 1 - \mathrm{HH}[Y_+(\cdot)]$ pode considerarse como indicador de mutabilidade, variabilidade ou diversificación relacionado co coeficiente presentado en Ramírez (1993). Tamén este indicador pode expresarse como unha norma $1 - HH = \left\|q(\cdot)\left(1(\cdot) - q(\cdot)\right)\right\|_1$, onde $1(\cdot)$ é a aplicación constante unidade. Pode interpretarse como unha media ponderada das co-cotas. A co-cota $1 - q(B_c)$ representa o que supoñen as outras clases no total global o que acrecenta máis a interpretación del como medida de mutabilidade ou diversificación. Así mesmo, pode verse como a diferenza entre os cadrados das normas 1 e 2 da partilla

$$1 - HH = \left(\frac{\|Y_+(\cdot)\|_1}{\|Y_+(\cdot)\|_1}\right)^2 - \left(\frac{\|Y_+(\cdot)\|_2}{\|Y_+(\cdot)\|_1}\right)^2 = \frac{\|Y_+(\cdot)\|_1^2 - \|Y_+(\cdot)\|_2^2}{\|Y_+(\cdot)\|_1^2} = \left\|\frac{Y_+(\cdot)}{\|Y_+(\cdot)\|_1}\right\|_1^2 - \left\|\frac{Y_+(\cdot)}{\|Y_+(\cdot)\|_1}\right\|_2^2$$
$$= \|q(\cdot)\|_1^2 - \|q(\cdot)\|_2^2.$$

No caso de que esteamos co recontador, se $C = 2$, $HH = f^2 + (1-f)^2 = 1 - 2f(1-f)$, polo tanto $1 - HH = 2f(1-f)$ onde $f$ é a frecuencia dunha das dúas clases.

A primacía é a cota máxima. No caso particular do recontador, a clase primacial é a moda e a primacía é o máximo da AF, a frecuencia máxima. Un posíbel indicador de representatividade da primacial é a súa co-cota que denominaremos a co-primacía. Adopta a forma

$$1 - \frac{\|Y_+(\cdot)\|_\infty}{\|Y_+(\cdot)\|_1} = \frac{\|Y_+(\cdot)\|_1 - \|Y_+(\cdot)\|_\infty}{\|Y_+(\cdot)\|_1},$$

Pode interpretarse como o que supón o resto das clases no total global ou o erro relativo de empregarmos o máximo no canto do total.

Outro indicador que pode interesar para dar unha orde de magnitude da $Y$ nas clases ou literais, un valor central, é $\|Y_+(\cdot)\|_1^{-1}\|Y_+(\cdot)\|_2^2$ porque

$$\min_c Y_+(\cdot) \le \frac{\|Y_+(\cdot)\|_2^2}{\|Y_+(\cdot)\|_1} \le \max_c Y_+(\cdot)$$

Ademais, verifica que

$$\frac{\|Y_+(\cdot)\|_1}{C} \leq \frac{\|Y_+(\cdot)\|_2^2}{\|Y_+(\cdot)\|_1},$$

considerándomos o produto escalar $<Y_+(\cdot)|1(\cdot)>$, que coincide con $\|Y_+(\cdot)\|_1$, e a desigualdade de Cauchy-Schwarz(-Bunyakovsky). Por iso, podemos interpretar a razón

$$\frac{\|Y_+(\cdot)\|_1^2}{\|Y_+(\cdot)\|_2^2}$$

como o número efectivo de categorías da nominal ou de clases da partición (*NEC*), dado que sería o cociente entre o total e un valor "medio".

Cando empregamos este indicador para compararmos dúas situacións no que o número de clases é diferente, $C \neq C'$, poderíamos recorrer a relativizar o indicador con respecto ao número, nominal, de clases

$$\frac{1}{C}\frac{\|Y_+(\cdot)\|_1^2}{\|Y_+(\cdot)\|_2^2}$$

e compararmos as dúas situacións mediante os seus indicadores relativos que denotaremos en minúsculas *nec*.

## 4. UTILIDADES

Seguindo unha orde determinada das clases, ben discrecionalmente, ben por tradición, por estándares ou por ela existir de xeito natural, acumulamos as cotas construíndo o que habemos denominar partilla acumulada; no caso particular das frecuencias, estamos perante aplicación de frecuencias acumuladas (AFA), $F(B_c) = F_c$ onde

$$F_c = \sum_{k=1}^{c} q(B_k) \ .$$

Con elas podemos construír dúas "funcións" graduadas 'escalonadas' que interpolan as seguintes secuencias de puntos

| | | | | | |
|---|---|---|---|---|---|
| Superior | $(0,0)$ | $(0,F_1)$ | $(F_1,F_1)$ | $(F_1,F_2)$ | $(F_2,F_2)$ … |
| Inferior | $(0,0)$ | $(F_1,0)$ | $(F_1,F_1)$ | $(F_2,F_1)$ | $(F_2,F_2)$ … |

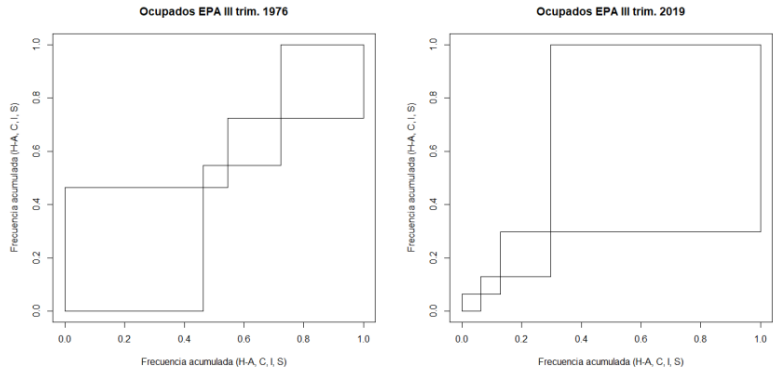
de xeito relativamente doado no R.

No primeiro exemplo que se presenta a seguir, as partillas acumuladas poden considerarse AFA, inda que haxa decimais son debidos á unidade que vén en millares. Como estamos a comparar dous trimestres diferentes cómpre manter a orde discrecional que neste caso é: halo-agrario, construción, industria e servizos.

**Ocupados por grandes sectores económico no III trimestre do ano correspondente.** **$10^3$ ocupado**

| Sector | 1976 | 2019 |
|---|---|---|
| **Halo-agrario** | 570,2 | 69,5 |
| **Construción** | 100,6 | 73,7 |
| **Industria** | 218,9 | 185,1 |
| **Servizos** | 338,8 | 777,4 |

Fonte: IGE Banco de series socioeconómicas de conxuntura




A suma das áreas dos cadrados representa o indicador $HH$. Por isto podemos denominalo diagrama dos cadrados diagonais. Observamos meirande concentración no ano 2019 que no 1976: HH1976= 0,329941 e HH2019= 0,5307454. Manifesta claramente a terciarización da economía que partía dun

estadio menos desenvolvido. O complementario da área encerrada polos cadrados case que representa o indicador $1 - HH$ agás que o cadro do gráfico é lixeiramente máis grande que o cadrado unidade.

De estudarmos o *NEC* nos dous anos, observamos que o NEC1976 é aproximadamente 3 (3,030845) mentres o NEC2019 é menor que 2 (1,884143). Outra mostra máis da terciarización coa perda de diversidade do aparato produtivo e a posíbel menor resiliencia ou enteireza del.

Nun segundo exemplo, o nomenclátor (IGE, 2019) amosaba que Vigo era o concello mais poboado de Galiza (293 642 hab en 2018 contra 244 850 da Coruña) mentres a entidade singular máis poboada era A Coruña (213 081 hab contra 199 598 de Vigo). A primacía da entidade singular (e.s.) da Coruña era 0,870251174 contra 0,679732463 da e.s. de Vigo.

No entanto, a superficie do concello de Vigo é meirande cá da Coruña (109,1 km$^2$ contra 37,8). Podemos considerar $R_1(\cdot)$, neste caso a densidade de poboación. $\|R_1(\cdot)\|_\infty$ valía 26 271,66311 na e.s. da Coruña fronte a 23 544,75778 da e.s. de Vigo e a razón global $R = \|Y_+(\cdot)\|_1 / \|X_+(\cdot)\|_1$, no concello da Coruña era 6472 hab/km$^2$ contra os 1248 do de Vigo. Como *HH*(Vigo) = 0,463608394, por ee.ss., mentres o da Coruña é 0,759032942, esta resulta máis concentrada. O número efectivo de ee.ss. é 1,317465876 na Coruña mentres en Vigo é 2,156992869. Se tivermos en conta o diferente número de ee.ss. destes concellos (Vigo: 301, A Coruña: 45), *nec*(Vigo) = 0,007166089 e o da Coruña, 0,029277019.

## 5. CONCLUSIÓNS

A perspectiva funcional –xa temos dito que as tabuladoras poden considerarse funcións de conxunto ou de argumento nominal– permite construír indicadores de xeito natural a partirmos das normas delas. As tabuladoras tamén serven, como paso intermedio, para construírmos diagramas a fin de visualizaren estes indicadores e mellorar a presentación dos datos.

Algúns deles xeneralizan indicadores xa coñecidos, os outros resultan máis novidosos. O *NEC*, entre os indicadores de diversificación presentados, está aliñado co *ID*. Este desconta as celas nulas mentres aquel tenta distinguir entre celas polo seu monto, máis ou menos apreciábel. Un é discreto e o outro continuo, polo que estende a súa aplicación. No entanto, a versión relativa daquel é máis difícil de interpretar como amosa o exemplo de que o concello máis poboado de Galiza é Vigo mentres a cidade máis poboada é A Coruña.

## REFERENCIAS

González Manteiga, W. e Iglesias Patiño, C.L. «Utilidad de los descriptores aplicativos y su estimación en la estadística oficial». *Estadística Española*, vol. 60, núm. 196, II cuatrimestre 2018. INE, Madrid

IGE (2019) Nomenclátor estatístico de Galicia. Ano 2018 en www.ige.eu

IGE (2019) Banco de series de conxuntura en www.ige.eu

IGE e DRN-INE (2000) *Comercio intracomunitario da Eurorrexión Galicia-Norte de Portugal. 1995-1997*, Compostela

Iglesias Patiño, C.L. (2019) «Os tabuladores, as súas compoñentes e mais a súa utilidade». *Actas do XIV Congreso Galego de Estatística e de Investigación de Operacións*, Vigo

Iglesias, C. e López, E. (2018) «Análise exploratoria da distribución espacial dos centenarios da Galiza». *Actas do III Encontro Luso-Galaico de Biometria*. Aveiro

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramírez Sobrino, J.N. (1993) *El análisis cuantitativo de la economía regional: los modelos econométricos regionales.* ETEA, Córdoba

U.S. Department of Justice (2021) «The Herfindahl-Hirschman Index» https://www.justice.gov/atr/herfindahl-hirschman-index

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# APLICACIÓNS DO MARCO INPUT-OUTPUT DE GALICIA: UNHA FERRAMENTA DE ESTATÍSTICA PÚBLICA PARA A ANÁLISE

Jácome Rodríguez, R.[1], Suárez Morao, M.[1] , Teijeiro Campo, M.T.[1], Calvo Ocampo, M.E.[1], Vilas Cruz, M.C.[1]

[1] Instituto Galego de Estatística

## RESUMO

O Marco Input-Output de Galicia (MIOGAL) é unha operación estatística elaborada polo Instituto Galego de Estatística (IGE). Desde finais do século XX, e en paralelo ao que ocorre nos países da nosa contorna, é unha ferramenta inserida na produción estatística oficial, e constitúe un dos piares do Sistema de Contas Económicas de Galicia. A súa capacidade descritiva da economía xunto coa súa vertente analítica achegan un abano de posibilidades que, en ocasións, están na fronteira das funcións dun instituto de estatística.

Nesta comunicación revisaremos as principais aplicacións do MIOGAL nos anos máis recentes desenvolvidas desde o IGE para dar unha maior difusión a esta operación pero tamén como apoio doutras iniciativas dentro da administración autonómica galega: a análise de sectores produtivos, as avaliacións de impacto macroeconómico ou o apoio a estudos concretos desenvolvidos na administración.

**Palabras e frases chave:** estatística pública, marco input-output, contas económicas, análise de sectores produtivos.

## 1. INTRODUCIÓN

O Marco Input-Output de Galicia (MIOGAL) é unha operación estatística consolidada cuxa elaboración é responsabilidade do Instituto Galego de Estatística (IGE). Ademais, é unha meta de información recollida no Plan Galego de Estatística, en concreto na sección "3.9. Sistema de Contas". O obxectivo central da operación é describir o proceso produtivo e os fluxos de bens e servizos da economía de Galicia. Outros obxectivos máis específicos:

- Afianzar o sistema de contas económicas de Galicia, definido como un marco contable integrado no que o MIOGAL é un piar no que se asenta a contabilidade anual e as contas trimestrais da economía galega.
- Realizar e dar solidez ás revisións extraordinarias regulares (*bechmark revision* ou *major revision*), anteriormente denominadas cambios de base contable. Estas revisións débense a cambios estatísticos, como a aparición de novas fontes de información ou melloras nos procesos de estimación, ou a cambios metodolóxicos introducidos por Eurostat. O propósito destes cambios é manter a "frescura" da información de síntese macroeconómico (IGE, 2019a).
- Facer un balance das fontes dispoñibles no sistema estatístico, dados os elevados requirimentos de información estatísticas necesarios para elaborar un Marco.

Na actualidade, o conxunto de táboas que forman o Marco está integrado dentro da produción estatística pública, e estreitamente relacionados coa elaboración da contabilidade nacional e rexional, pero non sempre foi así. Hai anos que Eurostat inclúe esta información entre o conxunto de datos que os Estados Membros deben remitir aos organismos comunitarios o que traslada a responsabilidade de elaborar esta ferramenta ás oficinas de estatística pública.

As aplicacións dun marco input-output teñen dúas vertentes moi diferentes: descritiva e analítica. As táboas elaboradas nesta operación proporcionan unha imaxe estruturada, simplificada e concreta dunha economía nun momento temporal dado, de aí a vertente descritiva. Pero tamén son a base da análise

input-output, que permite cuantificar sistematicamente as relacións entre os diferentes sectores da economía.

Galicia ten certa tradición na elaboración de táboas input-output. A primeira elaborada para a economía galega é previa á creación do IGE, foi elaborada polo *Servicio de Estudios del Banco de Bilbao* con ano de referencia 1980. O IGE participou na segunda edición (1990) e desde a difusión das Táboas Input-Output de Galicia 1998 todo o proceso de elaboración e difusión realízase integramente no instituto.

No mes de novembro de 2021 publicarase a actualización con ano de referencia 2018 do MIOGAL vixente (que ten ano de referencia 2016[1]).

Nos últimos anos, ademais de elaborar estas estatísticas o IGE difunde ou colabora na elaboración de análises aplicados con esta fonte de información. Nesta comunicación expóñense algunhas destas experiencias.

## 2. O MARCO NAS DIRECTRICES INTERNACIONAIS SOBRE CONTAS NACIONAIS E NA LEXISLACIÓN EUROPEA

Un usuario podería descargar a información do marco input-output dun Estado Membro da Unión Europea (UE) e comprobar como a información que alí se detalla é coherente coa información macroeconómica do seu país, que á súa vez serve de base para o deseño das políticas públicas da Unión. Isto é posible porque regulamentariamente a UE obriga aos Estados membros a remitir información dun xeito estruturado e sistemático.

A integración do marco dentro da produción estatística pública é un feito relativamente recente. No primeiro Sistema Nacional de Contas de Nacións Unidas (SCN-1953) non había referencia algunha ás táboas que forman o marco input-output, porén na seguinte edición (SCN-1968) mencionábase a integración do marco nun sistema de contas integrado. Desde a difusión do actual Sistema Nacional de Contas (SCN-2008) fortaleceuse a idea dun marco input-output como elemento integrador e que proporcione coherencia (United Nations, 2018).

En Europa, desde a aprobación do Sistema Europeo de Contas (SEC) de 1995, o marco input-output forma parte tanto da metodoloxía que debe seguirse na elaboración dos sistemas de contas como do programa de transmisión de datos que cada Estado Membro debe remitir a Eurostat.

O regulamento vixente (SEC-2010), que regula a idea descrita no primeiro parágrafo desta sección, indica que o núcleo central do marco input-output constitúeno as táboas de orixe e destino (TOD) a prezos correntes e a prezos do ano anterior. Tamén destaca que o marco se completa coas táboas input-output simétricas (TIO simétricas), derivadas das TOD a partir de certas hipóteses ou da utilización de datos adicionais (Eurostat, 2013). Nas TOD concorren as características máis descritivas do marco, mentres a parte analítica do mesmo reside nas TIO simétricas, como veremos a continuación.

A táboa ou matriz de orixe ofrece unha imaxe detallada da oferta de bens e servizos por produto e tipo de oferente, distinguindo no caso da oferta interior a produción por ramas, e no caso da oferta importada os bens e servizos adquiridos a unidades non residentes. É unha matriz rectangular, que amosa por columnas as ramas de actividade como produtoras no corpo central, e os vectores de importación, marxes comerciais e de transporte e impostos na parte dereita. Por filas aparecen os produtos da economía, polo que teremos información para cada produto da oferta total na economía, as orixes da mesma distinguindo no caso da produción interior as ramas que o producen e no caso da importación as orixes xeográficas (agrupadas en "resto de España", Unión Europea e "resto do Mundo").

A táboa ou matriz de destino informa sobre os usos da oferta de bens e servizos. Estes usos poden dividirse en consumos intermedios, é dicir, bens e servizos utilizados nos procesos de produción; consumo final (que pode ser dos fogares, das administracións públicas –AAPP- ou das institucións sen fin de lucro ao servizo dos fogares –ISFLSF-); formación bruta de capital (denominación do investimento no sistema de contas); e exportacións (vendas ao exterior). Ademais, na táboa de destino reflíctense as diferentes compoñentes do valor engadido por rama de actividade na economía. Tamén é unha matriz rectangular, cunha estrutura similar á anterior. No corpo central temos ramas en columnas e produtos en filas. As ramas aparecen aquí como consumidoras nos seus procesos produtivos (consumos intermedios) e

[1] Publicouse en decembro de 2019 cumprindo os estándares recomendados por Eurostat. Sendo t o ano de referencia, Eurostat recomenda a súa publicación dentro dun período de t+36 meses.

no resto da matriz, aparecen en columnas os usos da demanda final: consumo, formación bruta de capital e exportacións (por destino xeográfico, ao igual que as importacións).

Dise que son rectangulares porque o número de filas (produtos) é superior ao de columnas (ramas e resto de operacións). As ramas defínense consonte á Clasificación Nacional de Actividades Económicas (CNAE-09) e os produtos coa Clasificación de Produtos por Actividade (CPA), como así o indica o propio regulamento europeo (SEC-2010).

Na elaboración das TOD sintetízase moita información básica sobre cada unha das ramas da economía ou das operacións de bens e servizos que completan as matrices de orixe e destino. Neste aspecto reside a fortaleza descritiva, xa que é un exercicio de integración, contraste e equilibrio da información primaria dispoñible no sistema estatístico. Cada unidade institucional produtora de bens e servizos dunha economía clasifícase nunha rama de actividade en función da súa actividade principal e vaise rexistrar na táboa de orixe a súa produción ou producións (se ten varias) de cada un dos produtos descritos na matriz e na táboa de destino os bens e servizos que utilizou no seu proceso produtivo.

Existen dous tipos de identidades relacionando a información de ambas matrices:

- Identidade por rama: a produción de cada rama debe cubrir os insumos, é dicir, os consumos intermedios e as compoñentes do valor engadido bruto desde a perspectiva da renda: remuneración de asalariados, impostos netos sobre a produción e o excedente de explotación bruto.
- Identidade por produto: a oferta total por produto debe ser igual á demanda ou usos dos produtos. É dicir, para cada produto a produción máis as importacións son igual á suma de consumos intermedios, gasto en consumo final, formación bruta de capital e exportacións.

A táboa input-output simétrica é unha transformación das TOD, nas que se combina a información de oferta e demanda nunha única matriz. Esta combinación, realizada cunha serie de supostos, ten como obxectivo construír unha matriz que, no seu corpo central, conte co mesmo número de filas que de columnas, para aproveitar as propiedades alxebraicas e espremer as posibilidades analíticas das TIO simétricas. Nas filas e columnas desta xa non temos ramas e produtos senón ramas homoxéneas que producen un único produto. O obxectivo é reflectir un esquema de produción simple: cada rama de actividade só vai producir un único tipo de produción e farao cunha combinación de inputs: bens e servizos, factor traballo e capital[2].

As filas dunha matriz simétrica representan os destinos da oferta de cada rama homoxénea, que poden ser a demanda intermedia (resto de ramas de actividade homoxénea) e a demanda final. As columnas representan a función de produción de cada rama homoxénea. Para producir precísanse inputs (consumos intermedios), así como remunerar aos factores de produción: traballo (remuneración de asalariados) e capital (excedente bruto de explotación).

Os coeficientes input-output ($a_{ij}$) obtéñense dividindo cada valor da táboa simétrica, polo total da columna correspondente. Para estudiar as relaciones interindustriais nunha economía, traballaremos coa táboa simétrica da produción interior. É dicir, os $a_{ij}$ calculados son o resultado de calcular a produción da rama homoxénea i ($x_{ij}$) que é preciso para producir una unidade da rama homoxénea j ($X_j$).

$$a_{ij} = \frac{x_{ij}}{X_j} \qquad (1)$$

A produción de cada rama homoxénea pode ter como destino a demanda intermedia (input dun proceso de produción doutro ben ou servicio na economía) ou ben a demanda final (gasto en consumo, investimento ou exportación). Para *n* ramas homoxéneas o seguinte conxunto de ecuacións describirían o equilibrio oferta-demanda nunha matriz simétrica (2) :

$$\begin{gathered} x_{11} + x_{12} + \cdots + x_{1n} + D_1 = X_1 \\ x_{21} + x_{22} + \cdots + x_{2n} + D_2 = X_2 \\ \cdots\cdots \\ x_{n1} + x_{n2} + \cdots + x_{nn} + D_n = X_n \end{gathered}$$

[2] Mentres nas TOD os agregados por rama de actividade (a suma das columnas do corpo central das matrices) e os totais das operación de bens e servizos son os difundidos na contabilidade nacional ou rexional, nas TIO simétricas isto non ocorre, e só os totais da economía son coherentes con aqueles.

Substituíndo polos coeficientes input-output, a expresión quedaría do seguinte modo (3):

$$a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n + D_1 = X_1$$
$$a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n + D_2 = X_2$$
$$\cdots\cdots$$
$$a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n + D_n = X_n$$

En termos matriciais esta expresión é equivalente a:

$$AX + D = X \qquad (4)$$

Ou o que é o mesmo:

$$X - AX = D$$
$$(I - A)X = D$$
$$X = (I - A)^{-1} D \qquad (5)$$

Coñécese a expresión (I-A) como a matriz de Leontief e $(\mathrm{I} - \mathrm{A})^{-1}$ a súa inversa, sendo *I* a matriz identidad, *A* a matriz de coeficientes input output. Denominaremos matriz de multiplicadores técnicos interiores á inversa da matriz de Leontief (construída a partir da matriz simétrica da produción interior). Este é o modelo de demanda ou "static input-output model" (EUROSTAT, 2008), que calcula o impacto dun cambio na demanda (*D*) no total da produción da economía (*X*).

## 3. O MARCO INPUT-OUTPUT DA ECONOMÍA GALEGA

No último MIOGAL publicado a desagregación difundida foi de 72 ramas e 110 produtos nas TOD e de 71 ramas homoxéneas nas TIO simétricas (IGE, 2019b). A difusión por ramas das TOD é coherente coa difundida anualmente na operación estatística "Contas económicas anuais". A elección deste número de ramas débese a dous aspectos: dispoñibilidade de fontes estatísticas de base e relevancia para a economía galega de determinadas ramas de actividade. Con carácter xeral as ramas do MIOGAL están definidas polas divisións (dous díxitos) da CNAE-2009 porén, en determinados sectores, agregáronse ramas con escasa relevancia ou ben pola ausencia de información de calidade. Noutros casos, ofrécese un nivel de desagregación superior ás divisións, polo especial interese de determinadas actividades na nosa economía. É o caso da división 03, na que se distingue "Pesca" e "Acuicultura" e a división 10 "Industria da alimentación" para a que se distinguen ata cinco ramas de actividade diferentes. No que respecta aos produtos, cómprese o requisito detallado no SEC-2010, que indica que o número de produtos debe ser, cando menos, tan numeroso como o número de ramas.

Na información das TOD difundida polo IGE pode obterse información das producións e consumos intermedios por rama de actividade, e para estes, aqueles que son producidos na economía galega e aqueles que son importados doutras economías. Esta información define as relacións entre as industrias locais, e nela reside boa parte da análise input-output.

Para cada produto obtense información de oferta e información de demanda. Polo lado da oferta descríbese que parte da oferta provén de produción interior e que parte provén da importación. Polo lado da demanda descríbese como se utiliza da produto nos procesos produtivos doutras ramas de actividade (demanda intermedia) e como se utiliza na demanda final: canto se consume por parte dos fogares, AAPP ou ISFLSF, que parte da oferta ten como destino o investimento ou cal é a produción exportada pola economía galega.

Na difusión do MIOGAL tamén se inclúe a TIO simétrica. Como xa mencionamos a difusión realízase cunha desagregación de 71 ramas de actividade, e publícase información da TIO simétrica total, interior e importada. Nótese que, para avaliar os impactos e describir como se insiren cada un dos sectores (ramas homoxéneas) na economía galega utilízase a TIO simétrica interior, na que se inclúen só as compras que os sectores realizan dentro da economía galega.

A información contida nas TOD é coherente co resto de información dos sistemas de contas mentres as TIO simétricas son transformacións das TOD, baseadas en certos supostos, para obter unha estrutura coa que modelizar. En todo caso, os totais das operacións de bens e servizos (produción, consumos

intermedios, importacións, exportacións, gasto en consumo e formación bruta de capital) coinciden en ambos os dous conxunto de táboas, e son coherentes cos proporcionados nas contas económicas anuais.

Porén, na creación de ramas homoxéneas (produtoras dun único produto, para un determinado nivel de agregación dunha clasificación, e cunha única estrutura de inputs) a transferencia de producións secundarias acompáñase da correcta asignación dos consumos intermedios e os inputs primarios para desenvolver a dita produción. Para realizar correctamente esta asignación débese ter constancia dos factores produtivos utilizados en cada unha das producións secundarias das ramas de actividade da economía galega. Isto non pode ser afrontado como un exercicio de carácter estatístico, pola ausencia de fontes estatísticas que o permita, e pode considerarse como unha aproximación metodolóxica, xa que para a súa estimación apoiaranse nunha serie de hipóteses relativas á tecnoloxía utilizada para asignar correctamente os factores produtivos de cada produción secundaria (IGE 2019b)[3] Na figura 5 esquematizamos as relacións entre TOD, TIO simétrica e Contas económicas anuais.

Xunto coas matrices difundidas na TIO simétrica, calcúlase a matriz inversa de Leontief e acompáñanse dunha serie de cálculos derivados da análise input-output, en particular os coeficientes técnicos, os multiplicadores técnicos, os coeficientes de distribución e os coeficientes de traballo.

## 4. AS APLICACIÓNS: ANÁLISE DOS SECTORES SOBRANCEIROS DA ECONOMÍA GALEGA

Desde o ano 2012 consta na programación estatística a actividade de interese estatístico "Análise estatística de sectores produtivos e da estrutura económica en xeral". Baixo esta denominación o IGE publicou oito estudos sectoriais[4] nos que as estatísticas de síntese son a base da análise. Nestes documentos, ponse o foco nas columnas das TOD, é dicir, nas ramas de actividade que forman o sector e achégase tanto unha visión descritiva como outra máis analítica.

O último estudo difundido é a "Análise da Cadea Forestal-Madeira" que define este sector como agregación de catro ramas de actividade, é dicir, de catro vectores columna das matrices contidas tanto nas TOD como nas TIO simétricas:

- R02 Silvicultura e explotación forestal: inclúe actividades de cultivo de madeira en pé, explotación de viveiros forestais, a produción de madeira en rolla ou en bruto ou a recolección de produtos silvestres.
- R16 Industria da madeira e da cortiza, agás mobles: comprende a fabricación de produtos de madeira cos procesos de produción de serradura, cepilladura, conformación, laminación e ensamblaxe de produtos de madeira.
- R17 Industria de papel: inclúe a fabricación de pasta papeleira, papel e produtos de papel transformado.
- R31 Fabricación de mobles: esta división comprende a fabricación de mobles e produtos afíns de calquera material, agás pedra, formigón e cerámica.

Nas TOD os produtos relacionados coa cadea forestal madeira son seis: P02 Produtos e servizos forestais; P16A: Madeira serrada e cepillada; P16B: Outros produtos da madeira; P17A: Pasta de papel, papel e cartón; P17B: Artigos de papel e cartón e P31: Mobles.

Na vertente descritiva (visión das TOD) podemos ver como é o equilibrio oferta e demanda nos produtos da cadea forestal. Se agregamos os seis produtos (as seis filas das TOD) podemos construír un diagrama de fluxos no que combinemos a información de matriz de orixe e de destino, analizando tanto a composición da oferta como a dos usos dos produtos da cadea. Na parte esquerda da figura 1 aparece a fonte de oferta, en cor azul a produción galega e en cor verde as importacións, isto é, os produtos da cadea forestal que entran na nosa economía pero que foron producidos noutras economías. Cada un dos nodos que se aprecian na cor azul á esquerda son as ramas de actividade produtoras. Os módulos verdes responden a orixe da importación: resto de España, Unión Europea e resto do mundo.

[3] Na metodoloxía da operación estatística pode consultarse polo miúdo as hipóteses utilizadas para construír a TIO simétrica.

[4] Os sectores analizados foron: cadea forestal-madeira; automoción; pesca; téxtil, confección e calzado; cultura; transporte; minaría; agroalimentario.

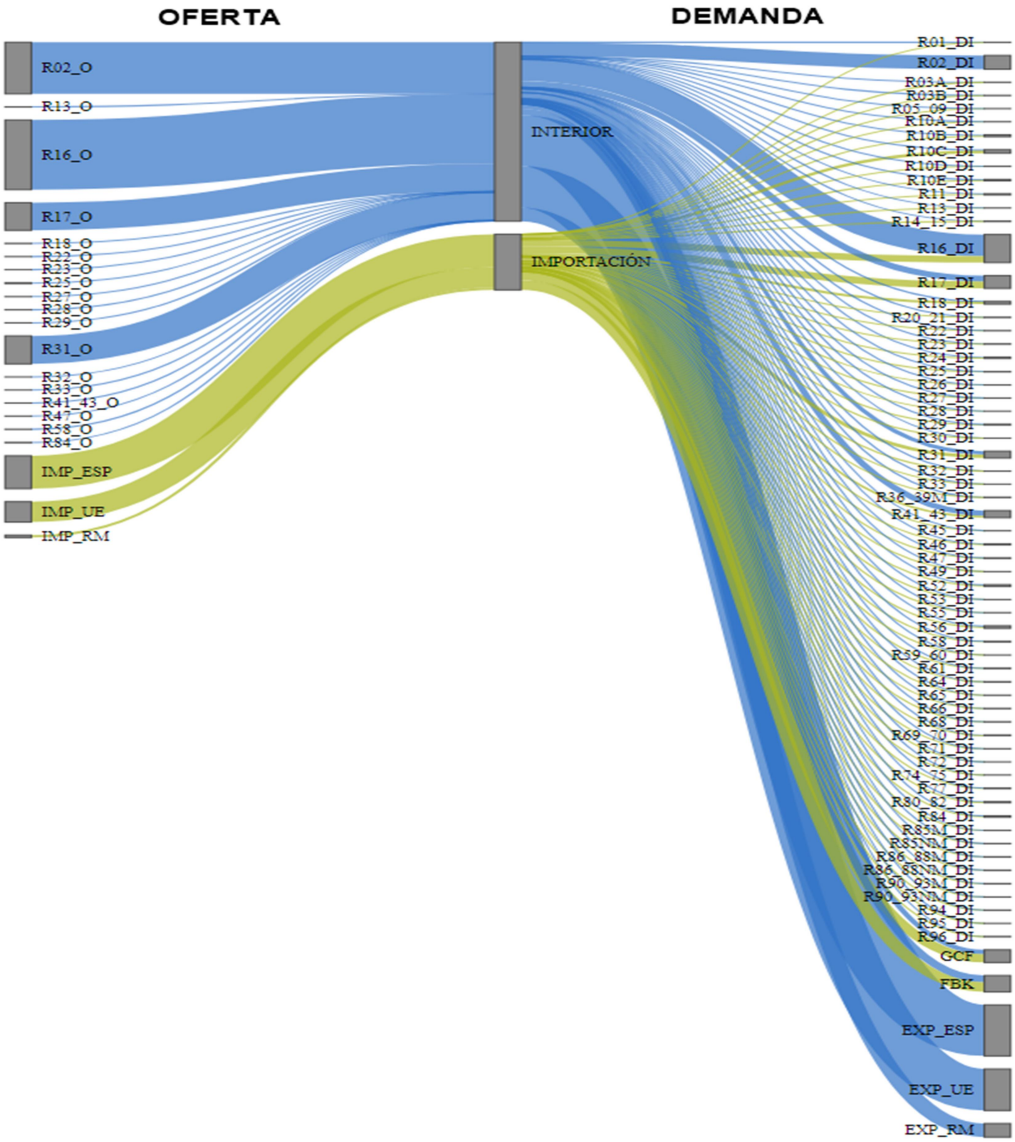



Figura 1: Diagrama de fluxos oferta-demanda dos produtos característicos da Cadea forestal-madeira. Fonte IGE. Marco Input-Output de Galicia 2016.

Na parte dereita poden verse as ramas consumidoras que son todas aquelas actividades que adquiren produtos da cadea forestal madeira nos seus procesos produtivos. E na parte inferior do gráfico vemos a produción que non se consume en Galicia e se exporta. Os produtos da cadea forestal son consumidos en moitas ramas de actividade, principalmente nas propias ramas de actividade que conforman a o sector, pero tamén en outras como a construción, por exemplo. Tamén cómpre destacar que hai unha parte relevante que é exportada.

Precisamente este é outro resultado que se pode extraer das TOD. Pódense calcular as porcentaxes de exportación de cada un dos produtos da cadea forestal-madeira, combinando información de oferta (produción) e de demanda (exportacións). Na figura 2 vemos os resultados: o 90% da produción de "17A Pasta de papel, papel e cartón" ten como destino os mercados exteriores. É a maior porcentaxe dentro dos produtos da cadea forestal-madeira que en agregado exporta aproximadamente o 60% da súa produción. A carón das tres cuartas partes da produción do produto "16B Outros produtos da madeira" ten como destino final a exportación.

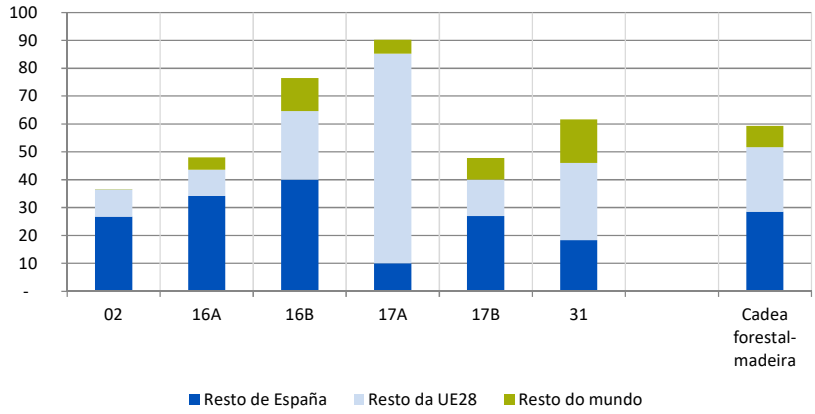



Figura 2: Porcentaxe de produción exportada. Fonte IGE. Marco Input-Output de Galicia 2016.

Na vertente analítica (TIO simétrica) nos estudos sectoriais que se publican no IGE céntrase a atención nas relacións entre sectores da economía galega fundamentalmente a través da análise input-output e con tres visións complementarias (IGE, 2021):

- Identificación de complexos produtivos: a través dos coeficientes de ligazón de Streit analízase como de estreitas son as relacións entre as ramas homoxéneas da TIO simétrica e como de extensas son esas relacións (número de ramas implicadas). Identificamos complexos produtivos se os coeficientes e o número de ramas relacionadas superan un limiar establecido a priori.
- Identificación de sectores clave: aqueles que son importantes demandantes a outras ramas de actividade e grandes oferentes de bens para outros sectores. A través da metodoloxía de análise input-output valóranse estas relacións como oferentes e demandantes e compáranse coas da media da economía galega.
- Análise de impacto: valora o que ocorrería se se incrementase nunha unidade a demanda final dun determinado produto (rama homoxénea). Este incremento faría que o sector produtor incrementase o seu output, para o que precisaría de novos inputs, que á súa vez deberían ser producidos por outros sectores da economía galega. A través da TIO simétrica e o modelo de Leontief pode valorarse o efecto final, derivado dun efecto directo (ou inicial) provocado polo aumento da demanda final e dun efecto indirecto provocado polas necesidades de novas producións a través dos sectores. Unha visión complementaria a esta son as simulacións do resultado de eliminar unha (ou varias) rama(s) sobre o conxunto da economía e os diferentes sectores produtivos, denominado método de extracción hipotética.

A modo de exemplo vemos a continuación a caracterización das ramas de actividade dos sector primario, industria e enerxético contido nos último análise sectorial (IGE, 2021). Esta análise, enfocada a avaliar a relevancia nas cadeas de valor das ramas forestal-madeireiras, indica que son sectores clave na economía galega a fabricación de vehículos automóbiles e compoñentes (R29), industria de procesado do pescado (R10B), enerxía eléctrica e gas (R35), a agricultura (R01) e a metalurxia (R24). Estes sectores caracterízanse por ser grandes oferentes de produtos utilizados nos procesos produtivos doutras ramas e ao mesmo tempo grandes demandantes de produtos producidos dentro da economía galega.

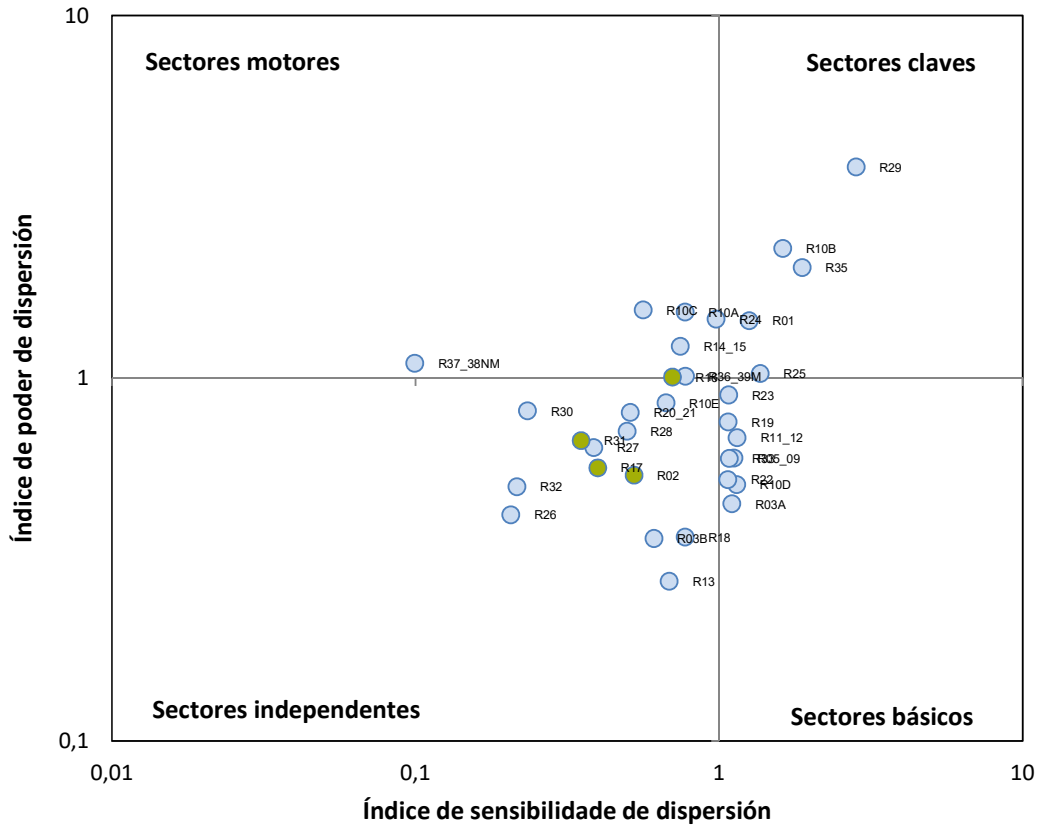



Figura 3: Caracterización de sectores produtivos (ramas do sector primario, industrial e enerxético). Fonte IGE (2021). "Análise da cadea forestal-madeira"

## 5. AS APLICACIÓNS: CONTRIBUCIÓN AO ANÁLISE DE POLÍTICAS PÚBLICAS

A Dirección Xeral de Planificación e Orzamentos da Consellería de Facenda solicitou ao IGE unha análise de impacto macroeconómico, isto é, unha cuantificación do efecto dos Fondos FEDER (para o período 2014-2020) no Produto Interior Bruto (PIB) e no emprego. Esta análise pode considerarse un primeiro paso de cara a establecer unha metodoloxía de avaliación do impacto na economía galega dos fondos xestionados pola Xunta de Galicia a partir do Programa Operativo FEDER-Galicia.

O estudo de impactos ante cambios na demanda final dunha economía é unha das aplicacións do "modelo de demanda" ou "static input-output model" (EUROSTAT, 2008) . Ante un cambio na demanda da economía, a análise input-output permite estimar o nivel de produción necesario para satisfacer o cambio, non só a produción necesaria nese sector no que existe unha nova demanda (efecto directo), senón o impacto naqueles sectores que producen inputs para aquel (efecto indirecto).

Neste caso, os fondos FEDER considéranse un estímulo da demanda final, sexa vía incrementos do investimento (formación bruta de capital) ou gasto en consumo final das administracións públicas.

Dado que o instrumento para realizar a análise está dispoñible no sistema estatístico (TIO simétrica), a fase crucial do traballo foi, a través da información detallada sobre os investimentos realizados e a actuación dos fondos, darlle o encaixe na metodoloxía SEC-2010 a cada importe do gasto rexistrado pola Consellería de Facenda desde dúas liñas:

- Identificar se se trata dun incremento do investimento, un maior gasto público ou se os fondos teñen outro tipo de destino.
- Identificar sobre que ramas homoxéneas redunda o maior gasto.

Con este análise previo, grazas ao detalle da información rexistrada pola Consellería e o asesoramento dos técnicos da Dirección Xeral de Planificación e Orzamentos, púidose construír un vector de impacto, é dicir, un vector que rexistre esa nova demanda na economía galega por rama homoxénea

derivada da dispoñibilidade dos Fondos FEDER en Galicia. A estimación do impacto no emprego ten un efecto vía consumo engadido aos efectos directos e indirectos, derivado do incremento de rendas salariais do novo emprego xerado polo impacto dos fondos.

Cando se produce unha nova demanda na economía hai que determinar a parte dela que vai ser satisfeita por produción interior. No caso que nos ocupa, unha liña de intervención consistía na adquisición de produtos tecnolóxicos, nos que Galicia ten unha produción moi escasa e a gran maioría provén da importación. Á hora de aplicar o vector de impacto hai que ter en conta este feito, polo que falamos neste traballo dun escenario de máximos, no que avaliamos o impacto na economía galega se toda a nova demanda fose satisfeita con produción interior e un escenario máis plausible no que estimamos a parte do gasto que vai ser satisfeita con produción das ramas de actividade galegas.

Os resultados indican un incremento do Produto interior bruto de Galicia no ano 2018 do 0,23% no escenario máis plausible. A meirande parte debido ao impacto directo (0,11%) e cun efecto indirecto (0,06%) similar ao efecto vía consumo (0,05%). No escenario máis plausible, un impacto de demanda de 120 millóns de euros provocan un efecto final de 145 millóns.

| | | Supostos de partida: | |
|---|---|---|---|
| | | Escenario de máximos | Impacto plausible da combinación de gasto público e privado asociado ao fondo |
| Anualidade cuberta por produción interior | | 181 | 120 |
| IMPACTO | Directo | 112 | 73 (0,11%) |
| | Indirecto | 53 | 41 (0,06%) |
| | Vía consumo | 46 | 31 (0,05%) |
| | Total | 212 | 145 |
| | %PIB | 0,33% | 0,23% |

Táboa 1: Impacto no PIB segundo supostos de partida. Ano de referencia 2018. Datos en millóns de euros. Fonte: Consellería de Facenda (2019)

Unha das fortalezas da análise baseada no input-output é a posibilidade de analizar impactos por ramas de actividade. No seguinte gráfico indícase o impacto no valor engadido xerado en cada sector, representado de forma agregada (10 agrupacións de ramas) unha información que está dispoñible para as 71 ramas homoxéneas da TIO simétrica.

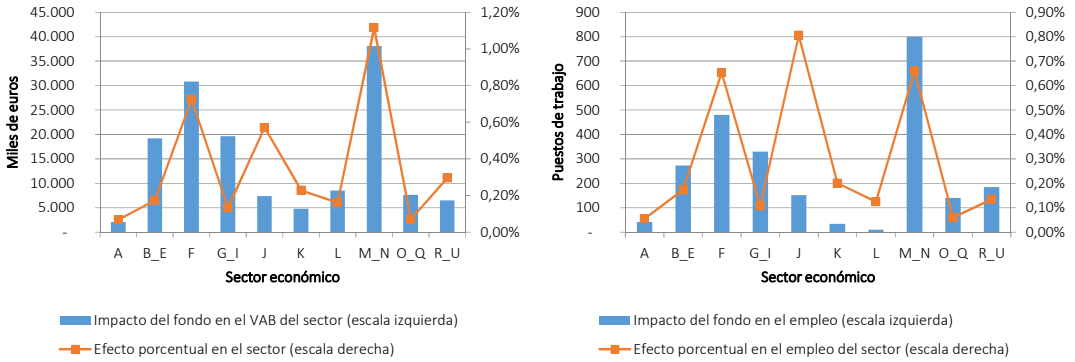



Figura 4: Impacto no PIB e emprego segundo supostos de partidas. Ano de referencia 2018. Datos do impacto no PIB en millóns de euros[5].
Fonte: Consellería de Facenda (2019)

[5] A Sector primario; B_E Industria e Enerxía; F Construción, G_I Comercio, transporte e hostalería; J Comunicacións; K Actividades financeiras; L Actividades inmobiliarias; M_N Actividades dos servizos a empresas, O_Q Administracións públicas, educación e sanidade; R_U Outros servizos.

Aínda que a metodoloxía seguida neste traballo é moi utilizada na análise de impacto de políticas públicas, non está exenta de debilidades. Trátase dun modelo de curto prazo, que descansa no suposto de que a estrutura económica non cambia e non é posible substituír factores produtivos entre si. O modelo non ten en conta que un grupo de medidas pode afectar dalgún xeito a outras medidas. É o que se coñece como hipótese de aditividade que no caso deste traballo viría a dicir que as liñas de axuda dos Fondos FEDER poderían afectarse entre si. Tampouco introduce o modelo corrección de prezos, nin ten en conta as expectativas dos axentes que participan na economía. Porén, entre as súas fortalezas está, como xa mencionamos, o elevado grao de desagregación sectorial que teñen, e neste caso, a relación con outras estatísticas do Sistema de Contas que permite facer unha análise complementaria de maior profundidade.

## 6. OUTRAS APLICACIÓNS

No ano 2020 o IGE colaborou co comité de expertos económicos de Galicia no estudo do "Impacto sectorial das medidas de distanciamento social"[6] debidas ao COVID-19. Membros deste comité propuxeron a utilización do Marco Input-Output de Galicia para adaptar un modelo deseñado polas Universidades de HEC (París) e Bocconi (Milán) que tiña por obxecto cuantificar o efecto das medidas de distanciamento social nos sectores produtivos da economía francesa (Barrot, Jean-Noel, et.al, 2020). A proposta do modelo e adaptación á economía galega foi unha proposta externa ao IGE, que xogou un papel de asesoramento no que respecta a información contida no MIOGAL ademais de proporcionar unha estimación da porcentaxe de empregos afectados polas medidas de distanciamento en cada rama homoxénea da TIO Simétrica.

O modelo proposto polos expertos parte deste vector de impacto que, a diferenza do modelo de demanda cuxas aplicacións se viron na sección anterior, resume as restricións da oferta de traballo nas primeiras semanas do confinamento o que provoca unha caída da produción e do valor engadido bruto xerado.

En moitas ocasións a desagregación dispoñible nas TOD e nas TIO simétrica non satisfai as necesidades dos usuarios que demandan unha análise máis afinada, ou cando menos unha desagregación sectorial máis ampla. Nun contexto de dispoñibilidade de información detallada poden construírse sub-vectores nas TIO simétrica para delimitar máis correctamente a incidencia dun subsector, empresa ou grupo de empresas. Para manter as propiedades e aplicacións input-ouput isto require equilibrar a nova matriz e realizar todos os cálculos de multiplicadores, etc. Esta é unha liña na que se está traballando desde o IGE.

## 7. CONCLUSIÓNS

Existe unha crecente utilización do Marco Input-Output tanto para aplicacións básicas, como o modelo de demanda, como doutras máis sofisticadas como a súa integración en modelos macroeconómicos máis complexos. Está consolidado o papel do Marco na produción de estatística pública tanto en Galicia como nas oficinas estatísticas da súa contorna e, no caso galego, estanse a facer avances desde o IGE nas aplicacións desta ferramenta analítica. Moitos destes avances proveñen das necesidades da propia administración autonómica e outros xorden da iniciativa do equipo que elabora esta operación estatística co obxecto de darlle unha mellor difusión á produción estatística. Este é o fin último desta comunicación ao XV Congreso Galego de Estatística e Investigación de Operacións.

[6] XUNTA DE GALICIA (2020), pax 99 e seguintes,

**REFERENCIAS**

Barrot, Jean-Noel, et al (2020). Sectoral Effects of Social Distancing. SSRN papers.

Consellería de Facenda (2019): Avaliación intermedia 2 do PO FEDER Galicia 2014-2020 (xuño 2019). Anexo I. http://www.conselleriadefacenda.es/documents/10433/3678822/Ev_Objet_2019_PO_FEDER_GA_14-20_AnexoI/3b05ddb7-eda6-4850-9dde-3cc0000c9542

EUROSTAT (2008): Eurostat Manual of Supply, Use and Input-Output Tables. Methodologies and working papers. Office for Official Publications of the European Communities

EUROSTAT (2013): *Sistema Europeo de Cuentas Nacionales y Regionales de la Unión Europea. SEC 2010*. Reglamento UE Nº549/2013 del Parlamento Europeo y del Consejo

IGE (2019a). Nota sobre a "revisión estatística 2019" nas operacións do Sistema de Contas Económicas de Galicia http://www.ige.eu/estatico/pdfs/s3/metodoloxias/Nota_Revision_Estatistica_2019_gl.pdf

IGE (2019b). Marco Input-Output de Galicia 2016. http://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0307007003

IGE (2021) "Análise da cadea forestal-madeira". Actividade de interese estatístico (AIE 13). Análise estatístico de sectores produtivos e da estrutura económica en xeral. http://www.ige.eu/estatico/pdfs/s3/publicaciones/AIE_Analise_Cadea_Forestal.pdf

United Nations (2018). Handbook on Suppy, Use and Input-Output Tables with Extensions and Applications. Department of Economic and Social Affairs. Statistics Division.

Xunta de Galicia (2020). Informe do comité de expertos económicos de Galicia para afrontar a crise provocada pola COVID-19. http://www.conselleriadefacenda.es/documents/10433/26507679/Informe-Completo-gal.pdf

**MATRIZ DE ORIXE**

| MATRIZ DE ORIXE | RAMAS DE ACTIVIDADE (CNAE) 1.....72 | Produción | Importacións | Oferta Total a Prezos Básicos | Marxes Comerciais e de Transporte | Impostos netos sobre os Produtos | Oferta Total a Prezos de Adquisición |
|---|---|---|---|---|---|---|---|
| PRODUTOS (CPA) 1 . . . 110 | Produción interior a prezos básicos | | | | | | |
| TOTAIS | | | | | | | |

**MATRIZ DE DESTINO**

| MATRIZ DE DESTINO A PREZOS BÁSICOS | RAMAS DE ACTIVIDADE (CNAE) 1.....72 | Demanda Intermedia | Demanda final | Demanda Total a Prezos Básicos |
|---|---|---|---|---|
| PRODUTOS (CPA) 1 . . . 110 | Consumos intermedios a prezos básicos | | Gasto en consumo, formación bruta de capital e exportacións | |
| TOTAIS | | | | |
| COMPOÑENTES DO VEB | Compoñentes do valor engadido por rama de actividade | | | |
| Remuneración de asalariados | | | | |
| produción | | | | |
| Excedente de explotación bruto | | | | |
| Postos de traballo (e PTE) | Postos de traballo (e PTE) por rama de actividade | | | |
| asalariados | | | | |
| non asalariados | | | | |

Datos dispoñibles na Contabilidade anual

Datos sen modificacións entre TOD e TIO Simétrica

Datos con modificacións entre TOD e TIO Simétrica

**MATRIZ SIMÉTRICA**

| MATRIZ SIMÉTRICA A PREZOS BÁSICOS | RAMAS HOMOXÉNEAS - PRODUTOS (CPA)- 1.....71 | Demanda Intermedia | Demanda final | Demanda Total a Prezos Básicos |
|---|---|---|---|---|
| RAMAS HOMOXÉNEAS - PRODUTOS (CPA) 1 . . . 71 | Consumos intermedios a prezos básicos produto por produto | | | |
| TOTAIS | CI totais a prezos básicos | | | |
| Impostos netos s/ produtos | | | | |
| CI a prezos adquisición | CI totais a prezos adqu. | | | |
| COMPOÑENTES DO VEB | Compoñentes do valor engadido por produto | | | |
| Remuneración de asalariados | | | | |
| produción | | | | |
| Excedente de explotación bruto | | | | |
| VEB | Valor engadido por produto | | | |
| Produción a prezos básicos | Produción por produto a pb | | | |
| Importacións | Importacións por produto | | | |
| Oferta total | Oferta a prezos básicos por produto | | | |
| Postos de traballo (e PTE) | Postos de traballo (e PTE) por rama homoxénea | | | |
| asalariados | | | | |
| non asalariados | | | | |

Figura 5: Relación entre TOD, TIO simétrica e as contas económicas anuais.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ENQUISA DE RESIDENTES EN GALICIA

Jaime Leirós Alonso de Velasco[1] e Rogelio López Romero[1]

[1] Área de Estudos e Investigación. Turismo de Galicia

## RESUMO

O desenvolvemento da actividade turística nun determinado territorio conleva a xeración dunha serie de impactos positivos e negativos en diferentes ámbitos. Algúns destes impactos poñen en cuestión a propia sostibilidade do turismo nun destino, motivo polo que están recollidos nas propostas de observatorio da sostibilidade como o da red INSTO da OMT. Vista a necesidade na nosa comunidade, a Área de Estudos e Investigación de Turismo de Galicia puxo en marcha no pasado ano 2020 unha enquisa dirixida aos residentes da comunidade para medir xustamente a súa percepción e valoración respecto da actividade turística e os seus múltiples impactos. Os resultados son aínda preliminares e xorden dunha análise puramente descritiva pero apuntan diferenzas relevantes na aceptación do turismo en función de determinadas variables sociodemográficas e de vinculación económica co sector.

**Palabras y frases clave:** Turismo, percepción, valoración, residentes, sostibilidade, xestión turística

## 1. INTRODUCIÓN

O desenvolvemento da actividade turística conleva necesariamente unha serie de impactos positivos e negativos —beneficios e custos— en diferentes ámbitos. Desta maneira podemos falar, con carácter xeral, de impactos ou consecuencias de índole económica —xeración de rendas, dotación de infraestruturas e servizos ou competencia por eles, incremento dos prezos...—, de corte social —interacción cos residentes, enriquecemento cultural, aversión aos turistas, perda de autenticidade...— ou medioambiental —coidado da paisaxe, xeración de ruídos, masificación e saturación...—.

Os impactos poden ser obxectivos ou subxectivos, sendo ambos de importancia capital como se aprecia no seguinte exemplo. Hipoteticamente, o turismo pode traer a un destino unha ganancia económica neta positiva, onde as rendas xeradas superan á presión sobre o coste da vida, e ao mesmo tempo pode causar a percepción contraria nos residentes, sendo fonte de aversión ao turismo e poñendo en cuestión a súa sostibilidade. Isto pode apuntar a problemas na distribución das rendas —obxectivos— ou a cuestións de novo máis subxectivas, pero ambos son decisivos e deben ser estudados e estimados mediante as técnicas de investigación social adecuadas.

Este é un exemplo, pero a sostibilidade do turismo como actividade económica e sociocultural depende dun amplo abanico de realidades. Para crecer en sostibilidade, para ser quen de avanzar na liña correcta, calquera que esta sexa, é preciso medilas e observalas. Turismo de Galicia, e en particular a súa Área de Estudos e Investigación, teñen en proxecto a creación de un Observatorio de Sostibilidade que integre todos estes aspectos para ser vector de cambio na realidade. Nesta liña de acción, xorde como primera gran carencia a solventar a falta de información sobre a percepción dos residentes galegos respecto da actividade turística no seu entorno, o que viña salientando a necesidade dunha nova operación estatística.

Existe unha extensa literatura científica que aborda esta necesidade específica, na que é posible ademais distinguir diversos enfoques e formas de abordaxe tanto no aspecto teórico como metodolóxico. Algúns deles presentan unha perspectiva analítica lineal mentres que outros, máis actuais, incorporan múltiples variables na análise. En calquera caso moitos deles parten da idea inicial de que a sociedade non é homoxénea nas súas características e tampouco nas súas percepcións e actitudes cara o turista, de aí que

sexa fundamental incorporar en calquera traballo a perspectiva segmentada, isto é, observar cambios de percepción na poboación residente en función de distintas variables socioeconómicas.

Desde a óptica da planificación e xestión turística pulsar o apoio da poboación resulta fundamental xa que a amabilidade e hospitalidade da xente é un elemento básico da experiencia turística, pero tamén de respaldo ás políticas públicas de desenvolvemento turístico. A nivel turístico, a apatía ou a desconfianza dos residentes terminan por afectar negativamente aos turistas e ás súa experiencia na viaxe, o mesmo que a boa acollida e a amabilidade redunda nunha relación positiva e mellora da propia experiencia. A nivel social, aproveitar a actividade turística como motor económico e social é un obxectivo desexable para a poboación residente, mentres que no límite oposto, antes de chegar a un hipotético rexeitamento maioritario na poboación, a actividade turística debe ser modulada.

A actitude, tal como se define polos investigadores, retroaliméntase de forma constante. Neste sentido un aspecto fundamental na análise de percepción dos residentes é a variable tempo, isto é, observar en que medida a percepción e valoración das persoas (neste caso a poboación residente) varía ao longo do tempo en función de diversos factores: o ciclo de vida do destino, o vínculo co sector, a situación económica e sociolaboral, etc.

Na análise da actitude dos residentes é importante ter en conta tamén, como factor condicionante, o perfil e as características do turismo e do turista que visita o destino. Distintos turistas (por perfil sociodemográfico, procedencia ou comportamento) xeran distintos impactos e, por tanto, distintas percepcións e actitudes dos residentes.

## 2. POSTA EN MARCHA E DESENVOLVEMENTO DA ENQUISA DE RESIDENTES

### 2.1. Antecedentes e xustificación do proxecto

O notable desenvolvemento da actividade turística a nivel mundial nos últimos anos, so interrompidos coa pandemia, ven provocando en paralelo un incremento da preocupación e malestar nos destinos máis maduros. Esta crecente aversión estendeuse rapidamente por destinos con niveis de afluencia moi afastados dos primeiros, pero que tomaron conciencia dos riscos e queren anticiparse a eles. En particular, no caso galego, a preocupación centrábase nos posibles niveis de saturación e de carga turística en determinados lugares como o Camiño de Santiago —nomeadamente o tramo final do Camiño Francés—, en Fisterra, no centro histórico da cidade de Compostela, e noutros destinos do litoral galego, por exemplo, sobre todo en épocas do ano con importante afluencia de peregrinos e visitantes.

Esta carga turística, aínda non medida nin avaliada en detalle no seu momento, pode xerar certo rexeitamento ou malestar en determinados segmentos da poboación local na medida en que entran en conflito os usos do espazo e de determinadas infraestruturas e servizos. Non en vano foron varias as novas en medios de comunicación que falaban dos perigos e amezas que en determinados lugares ou zonas de Galicia xurdían como consecuencia de posibles niveis de saturación turística.

Un dos escenarios onde repetidamente se alertaba dos perigos da afluencia masiva de peregrinos era o Camiño de Santiago. Con esta premisa, e intentando responder á necesidade de coñecer con profundidade os impactos que o Camiño ten a escala local, isto é, nos municipios por onde discorre, Turismo de Galicia puxo en marcha no ano 2017, en colaboración coa USC, un traballo dirixido precisamente a medir os impactos económicos, sociais e ambientais do Camiño Francés. Dito traballo incorporou na súa metodoloxía unha enquisa dirixida aos residentes dos municipios do camiño co fin de coñecer a súa percepción e valoración acerca do Camiño e das consecuencias que o desenvolvemento do mesmo tivo a diferentes niveis nos últimos anos.

Os resultados amosaron unha valoración global moi positiva, cunha percepción notable dos beneficios económicos do Camiño e certa neutralidade en canto aos impactos de índole social sen que se puidese confirmar a existencia de dificultades de mobilidade e de acceso aos servizos públicos por parte dos residentes nin tampouco unha actitude de rexeitamento ou aversión cara os peregrinos. A percepción global era, por tanto, positiva e mantívose un importante nivel de aceptación e de acollemento por parte da poboación local.

No canto dos turistas, neste caso os peregrinos, observouse un resultado sorprendente e moi relevante. A percepción de saturación é puramente subxectiva, relacionada coas propias expectativas e nomeadamente coa dinámica. Os peregrinos que perciben saturación son principalmente aqueles de largo recorrido que nos tramos finais atopan máis xente, incluso en épocas de pouca afluencia relativa. Pola contra, os peregrinos que transitan unicamente polos tramos máis concorridos e nos períodos de pico estacional, non teñen percepción de saturación.

Outros antecedentes sitúanse na cidade de Santiago de Compostela. No marco do Observatorio Turístico de Santiago realizáronse nos anos 2005 a 2007 varias enquisas dirixidas aos residentes na cidade para observar a súa percepción e valoración en relación ao desenvolvemento da actividade turística, principalmente no que atinxe ao centro histórico e contorno da Catedral.

Un dos factores que explican a importancia da poboación local no desenvolvemento do turismo nunha determinada zona é, como xa dixemos, o seu papel de "anfitrión" do destino. Neste sentido, diversos traballos e enquisas de medición da satisfacción do turista poñen de manifesto que un dos principais valores positivos de Galicia como destino turístico, ademais da natureza, a paisaxe e a gastronomía, é xustamente a hospitalidade dos residentes. Os turistas valoran de forma moi positiva, e mesmo o colocan como un atributo principal da nosa comunidade, o acollemento e a hospitalidade da xente. Dedúcese con isto que, con carácter xeral, o residente en contacto co turista ten unha actitude positiva fronte a actividade turística, sen ben non coñecemos con detalle se esta percepción é homoxénea ou varía en función da zona ou do momento do ano, por exemplo.

Esta inquedanza sempre estivo presente nos diferentes plans de desenvolvemento turístico levados a cabo en Galicia, e tamén dentro da órbita de Turismo de Galicia, na propia Área de Estudos e Investigacións. Por este motivo lanzouse no pasado ano 2020 a Enquisa de Residentes na comunidade como proxecto piloto que ten continuidade agora en 2021 e anos sucesivos.

Dous factores xogaron un papel desencadeante da posta en marcha da operación.

O primeiro deles ten que ver coa pandemia de COVID-19 e a situación de incerteza que se creou por mor do peche inicial e apertura gradual posterior das actividades non esenciais, entre elas o turismo, gravemente condicionado ademais polas restricións de mobilidade. Unha vez superado o pico da primeira onda da pandemia e coas primeiras medidas de desescalada, existía temor e incerteza na sociedade sobre os efectos que a mobilidade da poboación e a apertura da actividade turística podería ter na expansión da pandemia. O debate nos medios de comunicación e nas redes sociais era importante, e mesmo se coñecían casos de chegadas de visitantes a distintas zonas de Galicia, algúns de fóra da comunidade, para instalarse en segundas residencias ou en casa de familiares e/ou amigos, buscando fuxir de lugares de concentración de poboación ou de zonas con maiores restricións. Isto era un caldo de cultivo importante para o rexeitamento e fomentaba, sen dúbida, a denominada *turismofobia*, isto é, a aversión cara o visitante.

A circunstancia era excepcional e probablemente estaba moi afectada polo curto prazo e a rapidez nos cambios na normativa e nas restricións, pero ao mesmo tempo provocaba reaccións importantes que poderían repercutir e afectar a boa imaxe presentada ata o momento.

A esta preocupación engadiuse a necesidade do momento de coñecer, por parte dos xestores de turismo, o grao de interese dos mercados emisores por visitar Galicia no verán de 2020, unha vez rematado o período de desescalada. A ausencia de datos de fontes tradicionais, sobre todo na parte de predición, case obrigaba a lanzar unha enquisa propia para ver a intención de viaxe a Galicia e aproximar en que medida se poderían recuperar os niveis de demanda habituais na comunidade. As dificultades técnicas do momento e sobre todo a rapidez coa que sucedían os acontecementos levou a dirixir o traballo ao mercado interno —por ser o de maior peso na comunidade e tamén porque presentaba maior propensión a viaxar por Galicia: por ter menores restricións de mobilidade interna e polas dificultades e o temor de viaxar fóra da comunidade—.

O segundo factor detonante do proxecto ten que ver co marco conceptual e financeiro. Turismo de Galicia participa nun proxecto EP - INTERREG V A España Portugal (POCTEP), con outros socios da comunidade e do Norte de Portugal, denominado EURORREXIÓN DESTINO TURÍSTICO INTELIXENTE (en diante EDIT). O obxectivo do mesmo é a mellora da xestión dos recursos turísticos e a información sobre a actividade turística a través das TIC, transformando o modelo turístico en base á innovación, tecnoloxía, sostibilidade e accesibilidade de cara a incrementar a competitividade e rendibilidade do destino.

Inclúe varias liñas de acción entre as que se atopa a posta en marcha definitiva do citado Observatorio de Sostibilidade Turística, proxecto que contempla entre outras actividades a creación e desenvolvemento dun sistema de indicadores de medición de impacto da actividade turística, con datos obtidos de diferentes fontes. Unha delas, contempladas xa no propio EDIT, ten que ver coa medición da percepción e actitude dos residentes fronte a actividade turística e os turistas.

Con estas premisas no pasado ano 2020 a Área de Estudos e Investigacións puxo en marcha a Enquisa de Residentes de Galicia.

### 2.2. Desenvolvemento da enquisa: metodoloxía

No contexto do novo escenario turístico que xorde a raíz da pandemia, no que se observan notables incertezas entorno ao futuro máis inmediato e se producen cambios constantes nas pautas de comportamento da demanda turística e na percepción e valoración da poboación, a Área de Estudos e Investigacións de Turismo de Galicia puxo en marcha unha enquisa destinada aos residentes de Galicia cun dobre obxectivo:

- Coñecer a intención e perspectivas de viaxe no ano 2020, principalmente de cara ao verán, así como os gustos e preferencias
- Valorar a súa percepción como residentes acerca do turismo na nosa comunidade

Unha vez rematado o ano, e analizados os resultados, tomouse a decisión de dar continuidade á enquisa no ano 2021 —e sucesivos— pero centrando a mesma no apartado relativo á medición do sentimento e actitude dos residentes, modificando con isto os obxectivos e, por tanto, o cuestionario. Isto permitiu incorporar novas cuestións ao mesmo, ampliando a perspectiva de análise de percepción dos impactos.

Así, o cuestionario de 2020 incorporaba varios bloques de información:

a. Perfil sociodemográfico: lugar de residencia, xénero, idade, situación laboral, factores de risco COVID-19, beneficios do turismo
b. Perfil viaxeiro: viaxes por Galicia nos últimos anos
c. Intención de viaxe por Galicia no verán: con características da mesma en canto a duración, lugares de visita, tipo de aloxamento utilizado, etc
d. Valoración da presenza de turistas, nunha escala likert de 0-10

Esta última pregunta é central no cuestionario e de feito mantívose invariable en 2021 pois pivota sobre a mesma o eixo central de valoración do residente.

En 2021 o cuestionario cambia, elimínase toda a parte correspondente á intención da viaxe e céntrase exclusivamente en aspectos de percepción. Incorpórase un novo bloque de preguntas, cunha escala likert, de valoración sobre diferentes tipos de impactos do turismo na zona de residencia do enquisado pero buscando unha resposta balanceada sobre os aspectos positivos e negativos. Non se pregunta directamente por un impacto concreto senón pola posición do enquisado respecto dos factores positivos e negativos dun determinado tipo de impacto. Por exemplo:

a. O turismo contribúen á mellora do medio ambiente e conservación da paisaxe máis que á súa degradación
b. O turismo xera máis beneficios que prexuízos na miña localidade/municipio

O cambio de enfoque leva tamén a modificar as cuestións relacionadas co perfil sociodemográfico e laboral, así como calquera outro aspecto que permita realizar unha segmentación da poboación. Este proceso de mellora leva consigo tamén observar que variables son máis determinantes na confección do comportamento dos residentes e inflúen en boa medida na actitude dos mesmos. Desta maneira incorporáronse ao cuestionario cuestións como o nivel de estudos ou o detalle do perfil viaxeiro, e eliminouse a cuestión relativa aos factores de risco COVID-19.

O traballo de campo desenvolveuse en varios momentos do ano buscando observar en que medida a opinión e a valoración dos residentes cambiaba en función das circunstancias. Neste punto conflúen dous factores: un de carácter estrutural, isto é, a opinión dos residentes cambia en función da temporada turística polos distintos niveis de afluencia de visitantes —interese que permanece ao longo dos anos—; e outro de carácter conxuntural, no que tiña cabida a idea de medir a intención de viaxar.

No final da primavera de 2020 os criterios normativos e as restricións impostas cambiaban con elevada frecuencia, intentando obviamente adaptarse ás circunstancias e á evolución da pandemia. Neste sentido case dunha semana para outra as previsións de apertura de aloxamentos turísticos e de recuperación da actividade turística eran moi volátiles, cun elevado nivel de incerteza sobre o comportamento da demanda.

Por este motivo decidiuse dividir a toma de datos da enquisa en tres fases: unha previa ao verán —mediados do mes de xuño—, outra en plena temporada alta —finais do mes de xullo— e unha última despois do verán —segundo quincena do mes de setembro—. Desta maneira poderíase contrastar o carácter estacional das valoracións e da actitude dos residentes.

Este criterio mantívose en 2021 con leves cambios; adiántase a primeira onda da enquisa ao mes de maio como temporada media, divídese a onda de temporada alta en dúas fases, unha a mediados de xullo —para medir o inicio do pico— e outra nos primeiros días de setembro —para facer balance do verán

minimizando o nesgo de ter fora de casa aos residentes que viaxan— e retrásase a última onda ao mes de novembro como temporada baixa.

Este calendario está condicionado en boa media polo orzamento dispoñible e o reparto posible do número de enquisas ao longo do ano, buscando ter a representatividade suficiente en cada unha das fases. En 2022 está previsto incrementar o orzamento da operación e ampliar non só a mostra total senón tamén incorporar unha nova fase a comezos de ano, entre febreiro e marzo —antes de Semana Santa—.

O traballo de campo desenvólvese mediante unha enquisa telefónica dirixida aos residentes en Galicia maiores de 15 anos, utilizando o sistema CATI —entrevista asistida por ordenador—. Optouse por esta vía como a única posible no contexto de pandemia, pero xa estaba previsto manter esta estratexia debido aos menores costes que presenta fronte a alternativa de realizar a enquisa de forma presencial, nunha operación estatística que pretende representar correctamente un territorio amplo e heteroxéneo.

Emprégase unha mostraxe aleatoria simple estratificada. Defiríronse tres estratos segundo o concello de residencia, dividindo o territorio en tres grandes áreas: urbano —inclúe aos residentes nas 7 cidades de Galicia—, costa —residentes en concellos localizados no litoral galego— e interior —resto da comunidade—. Aplícanse posteriormente cotas de xénero e idade para manter unha correcta representatividade. A empresa responsable do traballo de campo realiza a mostraxe aleatoria sobre directorios de teléfonos fixos e móbiles pola súa conta.

O tamaño mostral fixouse en 2021 por motivos orzamentarios en un mínimo de 2.875 enquisas para o conxunto do ano. Tendo en conta a baixa no prezo por enquisa adxudicado, esta cantidade elevouse ata 3.396. Baseándonos nos datos de 2020, estimouse unha cuasivarianza de 7,2 na pregunta central de percepción do turismo receptor, de modo que con un tamaño mostral de 764 enquisas —cada unha das ondas que se fixeron este verán— obteríamos una precisión de 0,25 na media (valoración de 0 a 10), con una confianza do 99%. Para o conxunto do ano, temos un erro mostral teórico de 0,1185 na media da valoración con unha confianza do 99%.

Na primeira fase da enquisa, feita con moita urxencia en xuño de 2020, utilizouse un reparto territorial da mostra diferente. O criterio aplicado foi o de intensidade turística, medida esta como o cociente entre o número de prazas de aloxamento turístico e o volume total de residentes en cada un dos concellos de Galicia. A hipótese partía da idea de que a poboación residente nos municipios de intensidade alta teñen unha maior relación —directa e indirecta— coa actividade turística e, por tanto, a súa percepción pode ser diferente aos residentes en concellos con baixa intensidade turística. A aplicación deste criterio daba como resultado tres estratos, de intensidade alta —concellos con máis de 25 prazas por habitante—, media —concellos cunha ratio entre 6,5 e 25 prazas por habitante— e baixa —concellos con menos de 6,5 prazas por habitante—.

Este criterio de zonificación daba como resultado o seguinte reparto territorial:

| Estrato (Intensidade) | Nº concellos | Poboación |
|---|---|---|
| **Alta** | 27 | 101.965 |
| **Media** | 87 | 493.678 |
| **Baixa** | 199 | 1.783.765 |

Táboa 1: Zonificación segundo criterio de intensidade turística

O reparto proporcional da mostra prexudicaba claramente aos concellos de intensidade alta por ter un volume baixo de poboación. O criterio semella interesante, e unha primeira análise dos resultados así o indica, pero se observa unha notable heteroxeneidade territorial por canto, por exemplo, dentro dos concellos de intensidade alta están municipios de notable importancia turística como Sanxenxo e outros de baixo volume de poboación como Entrimo ou Triacastela. En cambio, no estrato de intensidade baixa colócanse, por exemplo, seis das sete cidades de Galicia —todas menos Santiago de Compostela— xunto con concellos de claro perfil rural e baixo volume de poboación.

As posibles dificultades de análise e un primeiro análise dos datos levaron a cambiar o criterio de reparto territorial por outro máis intuitivo, que ademais ofrecía perfiles diferenciados de percepción dos residentes. A zonificación quedou entón desta maneira.

| Estrato (Territorio) | Nº concellos | Poboación |
|---|---|---|
| **Cidade** | 7 | 995.126 |
| **Litoral** | 78 | 868.543 |
| **Interior** | 228 | 838.150 |

Táboa 2: Zonificación segundo criterio de localización

A mostra total da operación é de 2.649 enquisas no ano 2020 co seguinte reparto temporal:

| Período | Nº enquisas |
|---|---|
| **Xuño** | 519 |
| **Xullo** | 1.066 |
| **Setembro** | 1.064 |

Táboa 3: Reparto da mostra en 2020

En 2021 a mostra total ascende a 3.400 enquisas cun reparto temporal parello entre as catro ondas, das cales están executadas a día de hoxe as dúas primeiras, con 938 enquisas obtidas en maio e 764 enquisas en xullo.

Con ánimo experimental, na primeira onda de 2020 lanzouse, en paralelo á enquisa telefónica, un cuestionario online a través de diferentes vías: contas oficiais de Facebook e Instagram de Turismo de Galicia, e por correo electrónico á comunidade de seguidores do boletín de noticias e outras promocións e publicacións da axencia. O cuestionario foi desenvolvido na plataforma AGOL da Xunta de Galicia. Probouse este proxecto piloto coa idea de contrastar os datos coa enquisa telefónica e ver en que medida era posible obter resultados representativos por vías de menor custe económico, e pola oportunidade que ofrecía a plataforma en canto ás ferramentas de deseño e modificación do cuestionario e visualización a través de mapas e en tempo real do proceso de recollida de datos.

Os resultados foron analizados e amosan de inicio un claro nesgo cara a poboación de menor idade, de xénero feminino e cun perfil viaxeiro diferente por canto están máis vinculados —profesional e/ou persoalmente— ao turismo e presentan unha maior propensión a viaxar. Esta variable inflúe notablemente na percepción e valoración do turismo e, en consecuencia, nesga os resultados globais e incluso os específicos de cada segmento de xénero e idade.

| Estratos demográficos | Mostra Enquisa telefónica | Mostra Enquisa RRSS |
|---|---|---|
| **16-25 anos** | 9,1% | 19,0% |
| **26-35 anos** | 11,2% | 16,4% |
| **36-45 anos** | 17,3% | 23,9% |
| **46-55 anos** | 17,3% | 23,6% |
| **56-65 anos** | 15,6% | 13,8% |
| **66-75 anos** | 13,9% | 3,1% |
| **Máis de 75 anos** | 15,6% | 0,1% |
| **Home** | 48,4% | 28,6% |
| **Muller** | 51,6% | 71,4% |

Táboa 4: Comparativa das mostras telefónica e de RRSS. Xuño 2020

2.3. Principais resultados

A análise da enquisa fica de momento nunha fase puramente descritiva, da que se obteñen primeiros resultados de percepción e valoración da poboación residente con diferenzas de resultados para algúns segmentos concretos. En paralelo ao desenvolvemento do traballo de campo en 2021, e na medida en que

a mostra da enquisa aumenta, estase a traballar con unha análise cluster para detectar perfiles de residentes en función do seu posicionamento e valoración en relación á actividade turística.

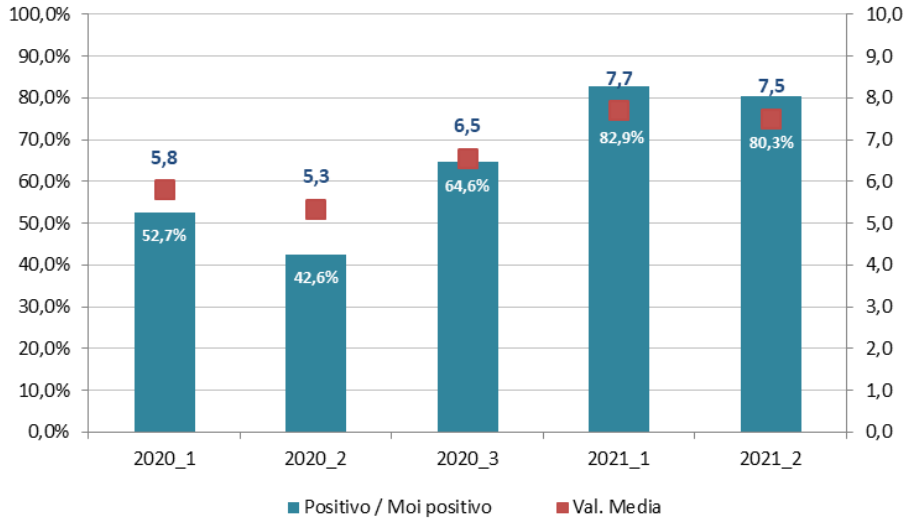



Figura 1: Porcentaxe de poboación receptiva á presenza de turistas

Unha primeira aproximación a partir da pregunta central do cuestionario amosa unha evolución e cambio de percepción da poboación neste último ano. O grao de acollida e a valoración positiva da presenza de turistas cambia en función do período do referencia pero xustamente neste momento de pandemia é complicado discernir que parte do cambio corresponde á temporada turística e que parte ten que ver coas características propias do momento, coa crise da COVID-19.

A primeira interpretación dos resultados amosa unha peor valoración e, por tanto, maior aversión á presenza de turistas en Galicia na primavera e sobre todo no verán de 2020 respecto de momentos posteriores. O enorme grao de incerteza que existía no momento da primeira desescalada da pandemia, e o temor a que a apertura da mobilidade no territorio nacional se traducise nun incremento dos contaxios e nunha nova onda da pandemia levou á poboación a mostrar maior receo á presenza de turistas. As porcentaxes de aceptación situábanse no 53% en xuño de 2020 e no 43% no mes de xullo, cunha valoración media de 5,3 nunha escala de 0-10.

Unha vez rematado o verán a percepción mellora e o grao de aversión diminúe, incrementándose a porcentaxe de poboación que considera positiva ou moi positiva a presenza de turistas en Galicia. Ascende ao 65% do total no mes de setembro e supera o 80% agora en 2021 (tanto en maio como en xullo), con valoracións medias que se sitúan en 7,7 e 7,5 respectivamente. Á espera de poder confirmar resultados en anos posteriores unha primeira hipótese relaciona a percepción da poboación coa evolución da pandemia.

Algunhas variables presentan certo grao de influencia sobre a percepción dos residentes. Así, por exemplo, un factor determinante é o vínculo da persoa co turismo en xeral, isto é, o grao de relación que ten coa actividade turística ben sexa desde o lado da oferta —traballador do sector, por exemplo, ou persoa que recibe algún tipo de beneficio do mesmo— ou da demanda —persoa que viaxa con certa frecuencia, por exemplo—. Neste sentido obsérvase unha relación directa entre o vínculo co sector e a percepción fronte ao turismo, sendo que as persoas que teñen algún tipo de beneficio —directo ou indirecto— presentan unha maior e mellor valoración fronte ao colectivo alleo á actividade turística.

Canto maior é o vínculo co sector, no sentido dos beneficios económicos obtidos de forma directa ou indirecta, mellor é a valoración e, por tanto, maior é o grao de aceptación. Isto leva a unha primeira conclusión: unha estratexia de socialización dos beneficios e de reparto territorial e demográfico dos mesmos levará a un maior grao de satisfacción da sociedade e, en consecuencia, a un maior apoio e aceptación da actividade turística.

Este elemento está moi relacionado cos beneficios de tipo económico que, tal como se verá posteriormente, teñen maior aceptación fronte a outro tipo de impactos como os socioculturais ou medioambientais.

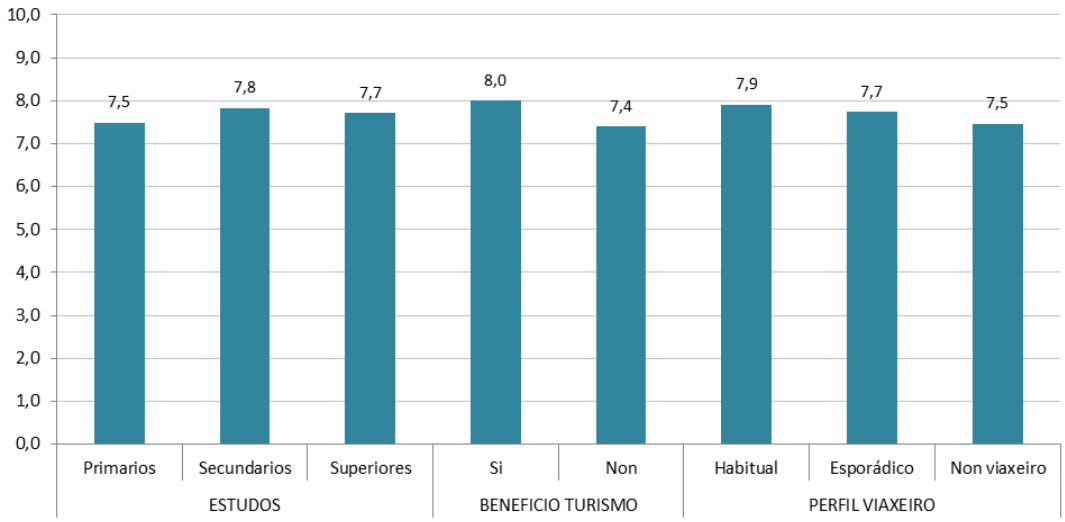



Figura 2: Valoración media da presenza de turistas segundo diferentes segmentos

A influenza do vínculo coa actividade turística tamén se aprecia desde a perspectiva da demanda, isto é, no papel do residente como turista ou viaxeiro. Neste sentido, apréciase unha mellor valoración naquel segmento da poboación que viaxa con frecuencia fronte a aqueles que, cando menos nos últimos anos, non realizaron ningunha viaxe turística, nomeadamente por motivos de ocio ou persoais. Os últimos resultados amosan unha maior aceptación por parte do viaxeiro habitual —definido como aquel que realizou un mínimo de catro viaxes no último ano—, fronte ao esporádico —aquel que realizou entre unha e tres viaxes nos últimos dous anos— e sobre todo fronte ao perfil non viaxeiro —que non realizou ningunha viaxe nos últimos anos—. As diferenzas non son en todo caso moi notables pero si se perciben matices entre eles, parece que de xeito consistente no tempo.

Esta mesma conclusión ten o seu traslado a nivel territorial. Aínda con datos moi incipientes e pendentes de contrastar e analizar con detalle nos vindeiros meses si se aprecia que este vínculo co sector ten unha influenza positiva desde o punto de vista territorial. Así, para os datos do mes de maio de 2021 obsérvase maior aceptación xustamente na poboación residente no grupo de concellos de intensidade turística alta fronte a aqueles de intensidade baixa. Tomando como referencia os valores promedio obtense unha puntuación de 8,1 nos concellos de intensidade alta, de 7,8 nos de intensidade media e de 7,4 naqueles de intensidade baixa. Non se aprecian neste caso diferenzas significativas na zonificación cidade-litoral-interior pero si unha mellor valoración no grupo de concellos que pertencen ao Camiño Francés —nos que varios presentan unha intensidade alta e se constata un notable peso do sector turístico—. A puntuación media na escala de 0-10 fica en 8,0.

Non se aprecian diferenzas significativas de valoración por grupos de idade —só cabe sinalar que nalgunha onda da enquisa a aversión ao turista semella máis alta no segmento de maior idade—, nin por xénero, nin tampouco pola situación laboral. En calquera caso esta conclusión extráese dunha primeira exploración descritiva e resta ver con análises máis profundas e detalladas o grao de influenza destas variables de tipo sociodemográfico.

O cambio de cuestionario en 2021 incorporou entre outros aspectos unha serie de cuestións relacionadas cos diferentes tipos de impacto do turismo. A valoración preséntase cunha visión integral, contrapoñendo os elementos positivos e negativos para obrigar dalgunha maneira ao enquisado a dar unha puntuación conxunta. Os resultados preliminares son os seguintes:

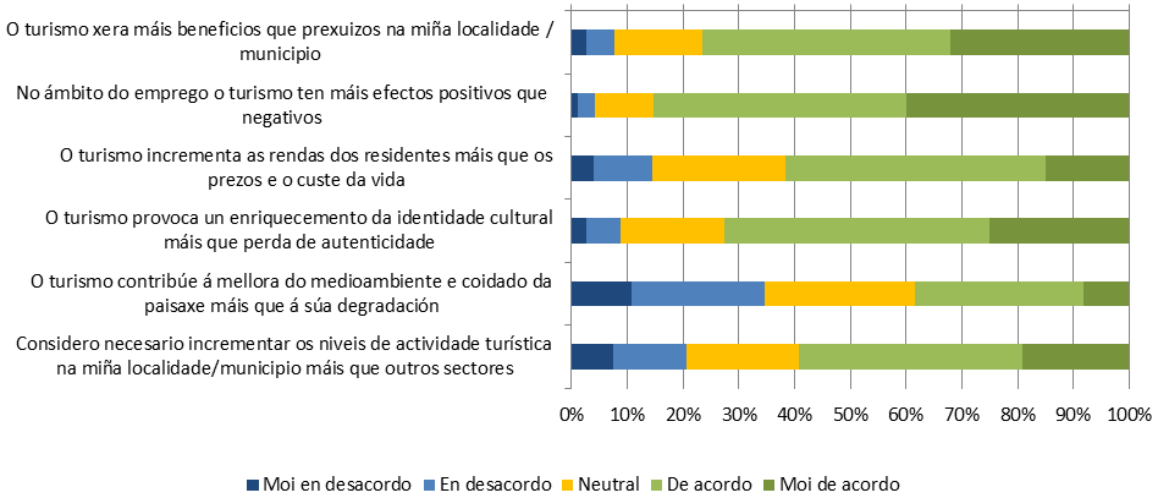



Figura 3: Opinión sobre diferentes impactos do turismo

En liñas xerais a poboación residente en Galicia ten unha valoración positiva sobre os efectos do turismo na comunidade. Máis do 50% está de acordo ou moi de acordo cos diferentes ítems presentados agás no caso dos impactos de índole medioambiental. Neste punto concreto a percepción é máis heteroxénea e case hai a mesma porcentaxe de poboación con opinión positiva e negativa. A valoración media é inferior ao resto do ítems o que indica unha maior preocupación por parte dos residentes na comunidade.

Os aspectos económicos do turismo —beneficios para o concello, efectos positivos no emprego— teñen unha mellor valoración e case se pode falar de certa unanimidade da poboación neste aspecto. Tres de cada catro residentes opina que o turismo xera máis beneficios que prexuízos na súa localidade, mentres que o 85% está de acordo en que a actividade turística ten máis efectos positivos que negativos sobre o emprego —valoran máis ou mellor as posibilidades de xeración de emprego sobre outros efectos negativos como a precariedade laboral ou os baixos salarios, por exemplo—. Os efectos en termos de obtención de rendas e incremento dos prezos, do custe da vida, son menos claros se ben segue sendo maioritario o volume de poboación que ten unha percepción global positiva.

| Ítems | Media | Desv. Típica | Coef. Variación | Beneficios SI | Beneficios NON |
|---|---|---|---|---|---|
| **Beneficios – prexuizos globais** | 7,4 | 2,28 | 0,31 | 7,8 | 7,0 |
| **Efectos positivos-negativos no emprego** | 7,8 | 1,95 | 0,25 | 8,2 | 7,5 |
| **Incremento rendas máis que custe vida** | 6,3 | 2,32 | 0,37 | 6,8 | 5,9 |
| **Impactos socioculturais** | 7,0 | 2,25 | 0,32 | 7,6 | 6,5 |
| **Efectos no medioambiente** | 5,0 | 2,57 | 0,51 | 5,3 | 4,7 |
| **Desenvolvemento turismo** | 6,1 | 2,68 | 0,44 | 6,4 | 5,9 |

Táboa 5: Efectos do turismo. Xullo 2021

Os aspectos de índole sociocultural teñen tamén unha percepción positiva superando o 70% a poboación que afirma estar de acordo ou moi de acordo coa idea de que o turismo provoca un enriquecemento da identidade cultural máis que perda de autenticidade. Non se aprecian neste sentido elementos negativos relevantes.

O aspecto que menos consenso suscita e menor valoración presenta ten que ver cos impactos de índole medioambiental. Ante a cuestión de se o turismo contribúe á mellora do medio ambiente e o coidado da paisaxe máis que á súa degradación os resultados amosan unha puntuación media inferior ao resto de ítems (5,0 na escala de 0-10) e unha maior heteroxeneidade —coeficiente de variación superior

ao resto—, e cunha proporción importante de residentes que opina a favor —o 38% do total— e en contra —o 35% neste caso—. Un 27% colócase nunha posición neutral —a porcentaxe máis elevada de todos os ítems considerados—.

Nunha perspectiva inversa podemos colocar as principais preocupacións dos residentes respecto dos efectos do turismo. Segundo os resultados preliminares da enquisa serían:

- As cuestións de índole medioambiental e o impacto da actividade turística no territorio e a paisaxe
- A capacidade do turismo para xerar niveis de desenvolvemento económico por encima doutros sectores ou segmentos de actividade
- Os efectos negativos do turismo sobre o incrementos dos prezos e do custe da vida

Tal como se observaba para os casos anteriores, o vínculo co sector —medido a partir da obtención de beneficios directos e/ou indirectos— ten unha influenza notoria na percepción dos residentes. Neste caso concreto as valoracións son máis elevadas para o segmento da poboación que percibe beneficios sobre aqueles alleos á actividade turística. Afecta ademais a todos os ítems con diversa intensidade. Tendo en conta os valores extremos, por exemplo, ascende a 8,2 a puntuación media sobre os efectos no emprego por parte do segmento que obtén beneficios e, no lado contrario, diminúe a 4,7 a valoración dos efectos no medio ambiente por parte do colectivo que non obtén beneficios. Este segmento alleo ao turismo amosa, en consecuencia, maior preocupación.

## 3. CONCLUSIÓNS

Os resultados preliminares da Enquisa de Residentes, obtidos dunha primeira análise descritiva pendentes de ampliar, profundar e mellorar con distintas técnicas estatísticas, amosan en liñas xerais un notable grao de apoio dos residentes en Galicia ao desenvolvemento da actividade turística na comunidade e unha importante acollida e actitude positiva fronte á presenza de turistas na súa localidade ou municipio.

Esta conclusión xeral presenta matices e diferenzas en función de diversos factores entre os que destacan o vínculo co sector, principalmente a capacidade ou oportunidade de obter beneficios directos ou indirectos do mesmo, máis tamén a relación estreita como usuario, isto é, como viaxeiro que tamén participa da actividade turística desde o lado da demanda.

Outros elementos de índole territorial —o grado de desenvolvemento turístico dun concello ou localidade, o nivel de intensidade turística, ou a localización litoral-interior— afectan tamén á percepción que o residente ten sobre o turismo e as características e dinámica do sector.

## REFERENCIAS

Cardona, J. R. (2012) Actitudes de los residentes hacia el turismo en destinos turísticos consolidados: el caso de Ibiza, (Tesis doctoral).

Cardona, J.R. e Serra, A. (2015) Segmentando residentes según sus actitudes: revisión de la literatura.

Gutiérrez, D. (2010) Las actitudes de los residentes ante el turismo (Tesis doctoral)

Gutiérrez Taño, D. e Díaz Armas, R.J. (2010) Las actitudes de los residentes hacia el turismo en un destino maduro. En Hernández Martín, R. e Santana Talavera, A. (coords) (2010) Destinos turísticos maduros ante el cambio: reflexiones desde Canarias pp. 255-280

Marzo-Navarro, M. (2017) Desarrollo del turismo rural integrado desde la perspectiva de los residentes: modelo propuesto

OMT (2007) Handbook on Tourism Market Segmentation

Rojas Tejada, A., Fernández Prados, J. e Pérez Meléndez, C. (1998) Investigar mediante encuestas: fundamentos teóricos y aspectos prácticos

Santos Peñas, J., Muñoz Alamillos, A., Juez Martel, P. e Cortiñas Vázquez, P. (2004) Diseño de encuestas para estudios de mercado: técnicas de muestreo y análisis multivariante

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# CONTROL DE CALIDADE DA INFORMACIÓN ESTATÍSTICA DIFUNDIDA POLO IGE

Mª Esther López Vizcaíno[1], Sergio Da Vila Davila[2] e Mª José Lombardía Cortiña[3]

[1] Instituto Galego de Estatística
[2] Universidade de Vigo
[3] CITIC, Universidade da Coruña

## RESUMO

Os institutos oficiais de estatística cada vez poñemos unha maior cantidade de información ao dispor do público. Este aumento na cantidade de información fai necesario o uso de ferramentas que permitan asegurar a calidade da información difundida. O obxectivo deste traballo é describir as ferramentas das que dispón o IGE para velar pola calidade da información dispoñible na web.

**Palabras e frases chave:** control de calidade, series temporais, atípicos, ARIMA.

## 1. INTRODUCIÓN

Ao longo das máis de dúas décadas de existencia, o Instituto Galego de Estatística (IGE) leva traballando para converterse nunha das principais fontes de información de datos de carácter socioeconómico da Comunidade Autónoma de Galicia e das súas principais divisións territoriais. Esta circunstancia motivou un incremento notable da cantidade de datos dispoñibles no Instituto, dispoñendo así, non só dos de produción propia, senón tamén dos de diversos organismos oficiais que ofrecen información estatística da nosa Comunidade Autónoma (Iglesias, 1999, 2001).

Ademais de aumentar o volume de información dispoñible na páxina web, nos últimos anos, púxose tamén énfase en mellorar a calidade, a accesibilidade e o intercambio da información estatística entre distintos usuarios e/ou plataformas. No IGE contamos cun banco de datos principal para difundir a información estatística dispoñible na nosa web.

A estrutura do banco de datos permite, ademais do acceso mediante táboas multidimensionais, que é a forma máis común de difusión de datos estatísticos na web do IGE, a captura directa de información por parte dos usuarios, o que evita a manipulación da información e facilita a actualización da mesma.

Durante os últimos anos incrementáronse de maneira exponencial os procedementos de carga automática da información que se difunde, co obxectivo de reducir custos e erros manuais. Esta baixa do número de erros manuais incrementou outro tipo de erros: de programación, de carga da información, etc. Isto fíxonos reflexionar sobre o reto de establecer procedementos que permitan avaliar a calidade da información publicada.

O obxectivo deste traballo é describir os procedementos de control de calidade da información publicada que están dispoñibles no IGE

## 2. METODOLOXÍA

En termos xerais, os controis que se realizan son os seguintes:

- Control de totais: que os datos sexan coherentes na dimensión do espazo (a suma de provincias debe ser o total, etc.) ou noutro tipo de dimensións (sexo, sectores, etc.).
- Coherencia do último dato: se é unha actualización de datos xa existentes, que o último dato publicado sexa coherente cos anteriores.
- Consistencia: Que as series de tempo que se publiquen sexan consistentes ao longo do tempo.

Os controis anteriores xunto coa metodoloxía para levalos a cabo explícanse nas seguintes seccións.

**Control de totais**

A práctica totalidade de toda a información estatística que se publica ten unha base territorial. No IGE podemos publicar información ata 6 ámbitos xeográficos nunha mesma táboa: Galicia, provincias, comarcas, municipios, distritos e seccións. Todos estes ámbitos xeográficos seguen unha xerarquía e é necesario asegurarse que a información que se publique cumpre esta xerarquía, é dicir que a suma das provincias é Galicia, que a suma das comarcas de Coruña é a provincia da Coruña, e así sucesivamente. Esta situación só a podemos utilizar cando o cálculo é para totais e non no caso de indicadores. O que é necesario que se cumpra para a dimensión do espazo, tamén o é para outro tipo de dimensións, como pode ser o sexo, é dicir a suma de homes e mulleres ten que ser o total, ou outros casos como a idade, a actividade económica, etc. Para facer o control de totais en todas as dimensións da táboa que se pretende publicar apoiámonos nunha librería do software R (R Core Team, 2018) desenvolvida no IGE e con nome `libhip` (Gómez, 2018) que permite, entre outras funcionalidades, ler as táboas multidimensionais e almacenalas en obxectos dunha clase S3 de R (`hip`); manipular a información contida na clase `hip` mediante operadores como filtrado, permutacións de eixos, …; crear táboas da clase `hip` e gravar táboas multidimensionais na base de datos.

Neste punto fixamos unha serie de conceptos que se repetirán ao longo do relatorio:

- Dimensións ou eixos: cada unha das características clasificatorias dos datos (sexo, idade, espazo, …)
- Membros: cada unha das modalidades que pode tomar unha dimensión: no caso do sexo, homes, mulleres e total.
- Niveis: as táboas multidimensionais poden ter unha xerarquía de membros presentando uns niveis segundo a profundidade nela: Galicia, provincia, comarca, municipio, son os niveis da dimensión espazo.

Aproveitando as funcionalidades que nos ofrece a librería `libhip` creamos unha nova función en R, con nome `controlTotal`, que avalía a coherencia da información en todas as dimensións e ao longo de todos os anos. Esta función terá como argumentos a táboa multidimensional, os números das dimensións dos eixos diferentes ao espazo onde se quere comprobar a coherencia e as relacións dos membros dentro dos eixos.

Por exemplo, coa táboa dispoñible na web do IGE sobre o *Paro rexistrado segundo xénero e grandes grupos de idade para Galicia, provincias, comarcas e municipios (https://www.ige.eu/igebdt/selector.jsp?COD=744&paxina=001&c=0201001002)* invócase a función creada coa seguinte sintaxe:

```
resultado= controlTotales(táboa, dimespacio,eixos, rel, pos, rele)
```

- *táboa*: é o código da táboa, neste caso a 744.
- *eixos*: dimensións (diferentes do espazo) onde se quere avaliar a coherencia, as dimensións 1 (Sexo) e 2 (Idade): c(1,2).
- *rel*: é necesario achegarlle as relacións dentro dos membros, por exemplo no Sexo o primeiro membro (Total) é a suma dos outros dous (Homes e Mulleres) e indícaselle que rel= c(1,2,2).
- *pos*: parámetro que indica os niveis que se quere comparar, neste caso o 1 co 2, c(1,2)
- *rele*: por último é necesario especificarlle as relacións entre os niveis do espazo rele= list( c(1,0), c(2,1), c(3,2)). Neste caso estáselle dicindo que Galicia é a suma de provincias, as comarcas suman as provincias e os municipios dentro de cada comarca, a comarca.

Esta función devolve unha lista de cada unha das dimensións cos posibles erros. Na Táboa 1 preséntase unha lista de erros para a dimensión do sexo. Nas tres primeiras columnas pódense observar as posibles combinacións de todas as dimensións excepto o sexo, e na última columna están as diferenzas (erros) entre o total e as sumas por sexo.

Táboa 1: Extracto de las diferencias existentes na dimensión do sexo entre as sumas dos membros do sexo (homes e mulleres) e os totais.

| Espazo | Idade | Tempo | Diferenza |
|---|---|---|---|
| Boimorto | Total | 2005 | 1 |
| Pino, O | Total | 2005 | -1 |
| A Barcala | Total | 2005 | 1 |
| Baña, A | Total | 2005 | 1 |
| Bergantiños | Total | 2005 | -1 |
| Coristanco | Total | 2005 | 1 |
| Laracha, A | Total | 2005 | -1 |
| Ponteceso | Total | 2005 | -1 |
| Betanzos | Total | 2005 | 1 |
| Aranga | Total | 2005 | 1 |
| Cesuras | Total | 2005 | 1 |
| Vilasantar | Total | 2005 | -1 |
| A Coruña | Total | 2005 | 2 |

**Coherencia do último dato**

Do mesmo xeito que no caso da dimensión do espazo, a práctica totalidade de toda a información estatística que se publica ten unha base temporal, son datos referidos a un intervalo de tempo e cunha periodicidade determinados: anual, trimestral ou mensual (na maior parte dos casos). Por tanto, como dimensión importante que é, é interesante estudala para controlar adecuadamente a calidade.

No IGE, e en xeral nos institutos de estatística, estase continuamente engadindo períodos de tempo á información dispoñible. Polo tanto, o primeiro control que teremos que facer é verificar que o último dato engadido é coherente cos datos dispoñibles para períodos anteriores. Nun primeiro momento propúxose calcular a taxa de variación (interanual, intertrimestral, intermensual,..) para todas as posibles combinacións e establecer un limiar que non debería superar. Isto ocasionounos bastantes problemas con datos inferiores a 100 unidades, por exemplo, que fluctúan moito e teñen taxas de variación que superaban os limiares considerados. Rexeitada esta posibilidade, botouse man da teoría das series de tempo. Polo tanto, a solución que se tomou foi:

- Axustar a serie temporal para todas as posibles combinacións dispoñibles na táboa, sen o último período, utilizando a metodoloxía ARIMA e os axustes automáticos que ten R. Utilízase o paquete seasonal (Sax, 2017) e a función de R `auto.arima` que busca o modelo ARIMA máis adecuado para a serie utilizando como criterio de selección o BIC (que penaliza o número de parámetros da serie).
- Calcular o intervalo de predición para o período seguinte (que é o que se vai a publicar).
- Comprobar que o dato que se vai a introducir está no intervalo de predición cun 95% de confianza.

Do mesmo xeito que para facer o control de totais, nesta ocasión tamén nos apoiamos na librería `libhip`. Neste caso construíuse unha función en R á que se invoca achegándolle dous argumentos, o código da táboa e o lugar que ocupa a dimensión do tempo:

```
resultado=coherenciaUltimoDato(1254, 1)
```

onde 1254 é o código da táboa, e 1 é o lugar que ocupa a dimensión do tempo.

A táboa 1254 ten información *sobre Emigracións, inmigracións e saldos migratorios para Galicia e provincias*(https://www.ige.eu/igebdt/selector.jsp?COD=1254&paxina=001&c=0201001002). Neste caso a función axusta 75 series de tempo. A saída desta función é un ficheiro coas series que teñen o último dato fóra do intervalo de predición, como se amosa na Táboa 2.

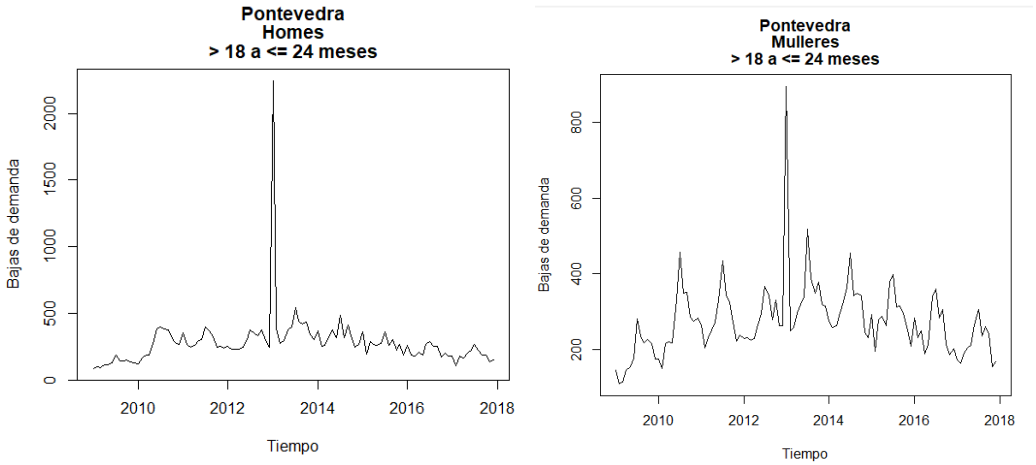
Táboa 2: Series cuxo último dato está fóra do intervalo de predición.

| | Espazo | Movementos.migratorios | dato | Point.Forecast | Lo.80 | Hi.80 | Lo.95 | Hi.95 |
|---|---|---|---|---|---|---|---|---|
| **311** | Galicia | Emigración externa | 21001 | 28853.00000 | 26032.7149 | 31673.2851 | 24539.7462 | 33166.25378 |
| **312** | A Coruña | Emigración externa | 9212 | 12549.20842 | 11401.7025 | 13696.7143 | 10794.2496 | 14304.16729 |
| **313** | Lugo | Emigración externa | 2681 | 3522.00000 | 3129.5716 | 3914.4284 | 2921.8325 | 4122.16752 |
| **314** | Ourense | Emigración externa | 2668 | 3915.00000 | 3343.6390 | 4486.3610 | 3041.1788 | 4788.82117 |
| **315** | Pontevedra | Emigración externa | 6440 | 11094.11528 | 10163.2454 | 12024.9852 | 9670.4727 | 12517.75790 |
| **466** | Galicia | A outra comunidade | 14862 | 19778.00000 | 17403.2376 | 22152.7624 | 16146.1144 | 23409.88561 |
| **467** | A Coruña | A outra comunidade | 6363 | 8248.00000 | 7333.6935 | 9162.3065 | 6849.6888 | 9646.31117 |

**Consistencia**

Neste punto partimos dunha situación onde xa temos unha cantidade importante de datos e táboas publicadas ás cales é necesario facerlles un control de consistencia de series, para detectar se hai datos erróneos. Por exemplo, na Figura 1 obsérvanse dúas series de tempo sobre as *Baixas de demanda de emprego para a provincia de Pontevedra para a duración da demanda de 18 a 24 meses e por sexo* e nelas preséntanse dous datos atípicos que, a priori, parecen erróneos.

Figura 1: Series de baixas de demanda de emprego para a provincia de Pontevedra para a duración da demanda de 18 a 24 meses



Polo tanto, como non partimos dunha situación inicial, é necesario validar a coherencia da información publicada nas táboas dispoñibles na web do IGE. Para isto, seguimos a metodoloxía de análise de series temporais que ten como obxectivo detectar valores atípicos da serie. Polo tanto, é de vital importancia ter unha ferramenta automática que permita a detección de valores atípicos en todas as series temporais dispoñibles na base de datos de difusión.

Un dos modelos máis empregados, que adoitan empregarse como punto de partida na análise de series temporais, son os modelos ARIMA, empregados tamén para a análise da coherencia. Os modelos ARIMA axustan os valores da serie tendo en conta as observacións anteriores e erros aleatorios cunha estrutura que permite incluír tanto compoñentes cíclicas como estacionais. Non obstante, a presenza de valores atípicos pode levar a unha incorrecta estimación dos parámetros do modelo, debido a que se poden ver nesgados polo efecto do atípico. Nesta dirección xurdiron métodos como o X-13ARIMA-SEATS e o TRAMO-SEATS que parten de modelos ARIMA, pero introduciron melloras como a detección e corrección de atípicos. O problema destes métodos é que están escalados para conxuntos de datos de tamaño pequeno e mediano. Para conxuntos de datos máis grandes xurdiu un novo enfoque para a detección de valores atípicos en series de tempo baseado en aplicar un proceso de descomposición e unha análise de residuos.

Este proceso en dous pasos baséase na idea de que visualizar valores atípicos en series temporais é complicado debido aos compoñentes estacionais e de tendencia. O primeiro paso aplica un modelo de descomposición que pode deixar estes dous compoñentes e extraer un residual. O segundo paso analizará este residuo, polo que o problema simplificaríase nun de detección de valores atípicos univariado.

Os métodos de descomposición considerados neste estudo son: STL, Twitter e STR. STL (Cleveland et al., 1990) foi deseñado coa idea de desenvolver un método sinxelo e rápido. Consiste nunha secuencia de operacións de suavizado realizado, todos menos un, polo mesmo suavizante: loess. Tamén inclúe unha versión robusta, que é interesante no caso de que o coñecemento previo indique que os datos teñen un comportamento non gaussiano. O método de descomposición STR (Dokumentov e Hyndman, 2015) foi desenvolvido para ser o marco máis xenérico para a descomposición de datos estacionais e para resolver algúns problemas que outros métodos non puideron. Algúns destes problemas son: incapacidade para proporcionar modelo estatístico sinxelo e manexable, incapacidade para calcular os intervalos de confianza ou incapacidade para ter en conta a estacionalidade múltiple. STR tamén ten unha versión robusta. E Twitter (Hochenbaum et ao., 2017) que é unha modificación de STL, na que a tendencia está representada pola mediana da serie de tempo.

Doutra banda, os métodos de detección de valores atípicos considerados neste estudo son: GESD, iForest e HDoutliers. Estes métodos foron elixidos entre todos os existentes debido ás seguintes caracteristicas. GESD (Rosner, 1983) é similar á proba de Grubbs ( Grubbs, 1950) pero aplícase nunha forma secuencial. É moi preciso cando o tamaño da mostra é n> 25. O principal problema é que, ao ser un procedemento iterativo, será máis custoso computacionalmente. iForest (Liu et al., 2009) manexa a detección de valores atípicos a través da idea de que, as observacións de valores atípicos representan a minoría dos datos e teñen atributos moi diferentes con respecto ás situacións non atípicas, polo que serán fáciles de illar do resto dos datos. Por último, HDoutliers ( Fraley e Wilkinson, 2020) baseou a súa idea de detección de valores atípicos na procura de espazos entre os datos ordenados, en lugar de só mirar os valores extremos.

Despois dun estudo de simulación levado a cabo cun conxunto de series representativo das publicadas no IGE decidiuse aplicar a seguinte combinación de métodos para a detección de outliers:

- TRAMO-SEATS
- STL+iForest
- STR+HDoutliers
- STR+iForest

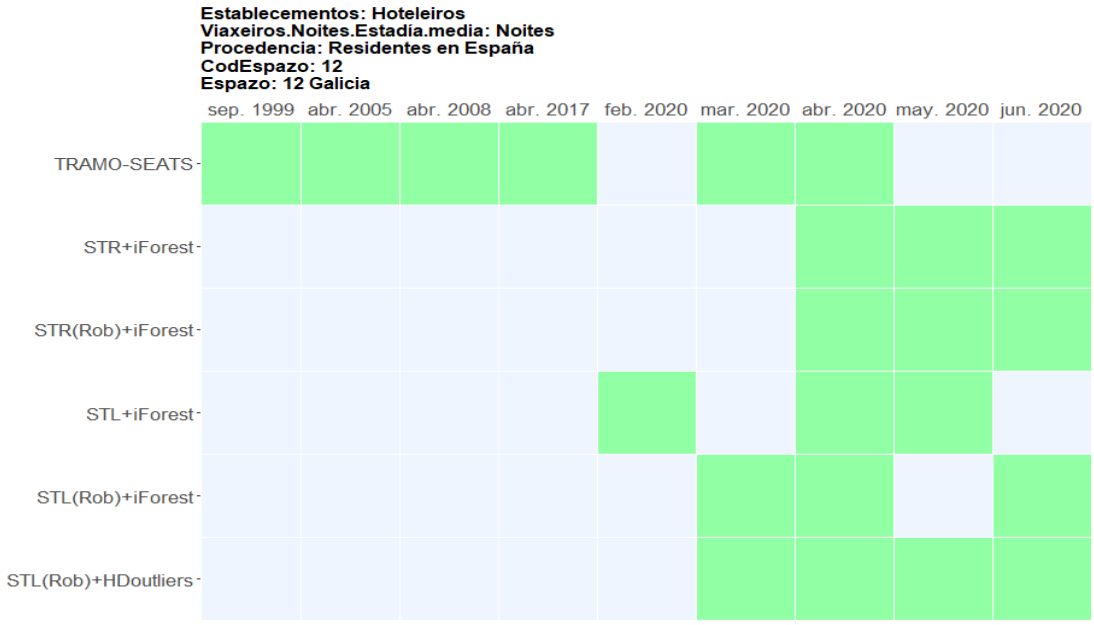
Nesta situación construíuse unha función de R, con nome `consistencia`, á que se lle invoca achegándolle dous argumentos, o código da táboa e o lugar que ocupa a dimensión do tempo.

```
resultado=consistencia(tabla,dimtempo)
```

A saída desta función é un ficheiro coas series que teñen atípicos e un diagrama de calor para cada unha destas series. Na cabeceira deste diagrama de calor reflíctese a serie que ten atípicos, nas columnas

os períodos atípicos e nas filas os métodos empregados. Un exemplo de diagrama de calor móstrase na Figura 2.

Figura 2: Diagrama de calor da serie de Pernoitas de viaxeiros residentes en España e que visitan Galicia. Nas filas os métodos empregados e nas columnas os períodos candidatos a atípicos.




Nota: (Rob) versión robusta

O criterio empregado para que un período se considere atípico é que todos os métodos empregados débeno marcar como atípico.

## 3. CONCLUSIÓNS

Neste traballo abórdase o problema da importancia de efectuar un bo control de calidade sobre a información que se difunde nun instituto de estatística. Descríbense varias ferramentas que hoxe en día se están utilizando no IGE.

Desde que se empezou a utilizar estas ferramentas detectáronse erros de control de totais, de consistencia e tamén na inclusión do último dato, que tiñan a súa orixe, tanto en erros manuais de carga de información, como de fallos nos programas de automatización da carga da información.

Todos estes erros detectados reforzan a idea de que é necesario realizar un control de calidade da información antes de publicala que asegure que os datos que se están publicando son correctos.

## REFERENCIAS

Cleveland, R. B., Cleveland, W. S., McRae, J. E., y Terpenning, I. (1990, 01). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6.

Dokumentov, A., y Hyndman, R. J. (2015, 01). STR: A seasonal-trend decomposition procedure based on Regression.

Fraley, C., y Wilkinson, L. (2020). Hdoutliers: Leland wilkinson's algorithm for detecting multidimensionaloutliers [Manual de software informático].

Gómez, J. (2018). Tablas multidimensionales en R. XX Jornadas de Estadística de las Comunidades Autónomas. Logroño.

Grubbs, F. (1950, 03). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21 . doi: 10.1214/aoms/1177729885.

Hochenbaum, J., Vallis, O., y Kejariwal, A. (2017, 04). Automatic anomaly detection in the cloud via statistical learning.

Iglesias, C. L y Arias, A. (1999). A aplicación de consulta de series no web do IGE: un instrumento para o estudio da conxuntura económica da C. A de Galicia. *Boletín de Series estatísticas de Galicia* Nº 47.

Iglesias, C. L. (2001). La nueva política de Difusión del Instituto Galego de Estatística. Jornades europees d'estadística. Palma.

Liu, F. T., Ting, K., y Zhou, Z.-H. (2009, 01). Isolation forest. En (p. 413 - 422). doi: 10.1109/ICDM.2008.17

Rosner, B. (1983, 05). Percentagepoints for a generalized esd many-outier procedure. *Technometrics*, 25 , 165-172. doi: 10.1080/00401706.1983.10487848

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sax, C. (2017). seasonal: R Interface to X-13-ARIMA-SEATS. R package version 1.6.1. https://CRAN.R-project.org/package=seasonal

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Análise estatística multivariada das comunidades intermunicipais de Portugal

Miguel Marques de Sousa[1], António Blazquez Zaballos[2]

[1] Doutorando do Departamento de Estadística de la Universidad de Salamanca
[2] Profesor en el Departamento de Estadística de la Universidad de Salamanca

## RESUMO

Tomando como fulcro a Região Norte de Portugal, pretendemos fazer uma caracterização inter-regional do país, utilizando as técnicas da análise estatística multivariada.

Até à segunda Grande Guerra, o território nacional, apresentava um dinamismo económico e social, característico do subdesenvolvimento, mais ou menos homogéneo de Norte a Sul e de Este a Oeste. Salvaguardando as características específicas das regiões urbanas e rurais e de interior e litoral, não se verificavam nessa época grandes assimetrias regionais, excetuando as resultantes, "naturalmente", dos polos de desenvolvimento que constituíam as grandes cidades de Lisboa e Porto.

De então para cá, assistiu-se progressivamente à drenagem de gentes e atividades económicas, que reduziram o interior, especialmente a periferia, ao "deserto" que é hoje. De pouco têm valido as políticas económicas tendentes a contrariar o processo de despojamento do interior.

Para o efeito, utilizando o *software* R, aplicou-se a Análise de Componentes Principais (ACP) aos dados do censo de 2011, das 23 sub-regiões de Portugal continental, subentendidas como comunidades intermunicipais. Cada sub-região é uma unidade estatística, observada sob um conjunto de 19 variáveis, das áreas económica, demográfica, social e política.

Sendo o objetivo classificar as variáveis e respetivo grau de influência que exercem, quer nas restantes variáveis quer nas unidades estatísticas, verificou-se que são as variáveis populacionais e as económicas, as que maior interação têm na segmentação das comunidades intermunicipais, entre litoral e interior e no processo bipolar em torno das duas maiores áreas metropolitanas do país. Conclusivamente, as políticas socioeconómicas aplicadas por sucessivos governos durante as últimas sete décadas, nunca conseguiram atrair investimento económico e, obviamente, fixar populações nas sub-regiões mais periféricas de cada um dos dois polos do país.

**Palavras e frases chave:** *População; economia; interagir; comunidade intermunicipal, periferia*.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ANALYSIS OF THE OVERALL ISONYMY IN GALICIA UNDER A COOPERATIVE GAME THEORY APPROACH

Mª José Ginzo Villamayor[1] and Alejandro Saavedra Nieves[1]

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

## ABSTRACT

The analysis of patterns in the distribution of surnames provides information on population movements and its characteristics. Isonymy measures the regional distribution of surnames in different geographical areas or historical periods. The main goal of this work is the usage of specific solutions of transferable utility games, such as the Shapley value, as a tool to measure the marginal influence of each region on isonymy, under the idea of their merging. For this purpose, the concept of isonymy is extended for the case of more than two regions in an innovative way. This proposal is applied on the data of the Galician surnames for the census data of 2011. From the obtained results, some pattern and distribution of the different types of surnames in Galicia are inferred.

**Keywords**: Surnames, isonymy, isonymy between, Lasker distance, Shapley value, ranking.

## 1. INTRODUCTION

The study of surnames was traditionally used for the knowledge of the genetic structure of populations and is currently being used as a source of information to characterise the population of a region. The analysis of the patterns observed in the distribution of surnames provides information on movements and characteristics of populations and consequently allows the study of relationships with other factors such as history, language, genetics, demography, among others. There are many precedent works that use isonymy (i.e. the possession of the same surname) to analyse the regional distribution of surnames in different geographical areas and also in different historical periods. For the case of United Kingdom, Cheshire et al. (2010) showed a strong relationship between surname regions and geographical regions. Boattini et al. (2012, 2010) analysed the usefulness of surnames for understanding the ancient history of the populations of the Italian peninsula. Using multivariate methods, Novotny and Cheshire (2012) identified a clear parallelism between surname regions and ethnocultural boundaries in Czech Republic. Similar studies were carried out in Mikerezi et al. (2013) in smaller geographical spaces, such as Albania. Ginzo-Villamayor et al. (2013) also show how the study of surnames through isonymy measures combined with cluster analysis tools can provide insight into the urbanisation processes in rural and urban areas of Galicia. Despite the extensive literature on the identification of surname regions through isonymy measures, no relevant methodological advances have been observed in recent years and the analysis procedures used are essentially the same (*Lasker*, *Nei*, *isonymy between areas*, etc.), except for some exceptional work such as that of Boattini et al. (2012).

In this paper, we present an innovative approach to rank regions according to the marginal influence on the overall isonymy by assuming the hypothesis of their merging. To this aim, we use some models from cooperative game theory and, in particular, specific solutions of transferable utility games as the Shapley value (Shapley, 1953). To study the isonymy structure of Galician councils, the surnames of the 2011 census in 315 councils in this region were analysed. The study was carried out in two parts. First, we introduce different measures of isonymy between regions, extending the well-known notions of isonymy between or Lasker distance to the case of considering more than two areas. This fact allows the definition of two new classes of TU-games related to isonymy settings. After, we apply our proposal for ranking the Galician councils using as criterion their influence on the overall isonymy.

## 2. BASIC NOTIONS ON ISONYMY

Surname (dis)similarity between regions can be quantified by different measures. Let $R = \{1, \ldots, r\}$ be the set of $r$ geographical regions. Each region $i \in R$ has associated a collection $S_i$ of surnames.

Isonymy (Lasker, 1968) for a certain region $i \in R$, denoted by $I_i$, refers to the possession of the same surname, so isonymy can be connected to biological relation, being a premise in genetics that individuals with the same surname are more likely to share the same family lineage. Formally, a measure of isonymy in a region it is given by

$$I_i = \sum_{l \in S_i} p_{li}^2, \text{ for all } i \in R, \tag{1}$$

where $p_{li}$ denotes the relative frequency of surname $l$ in region $i$. High values of isonymy are possible in a population where there are relatively few surnames, and low values of isonymy are obtained when the number of surnames is relatively large.

As notation, $(i, j)$ is the pair of regions $i$ and $j$, with $i, j \in R$. Thus, the joint collection of all the surnames of $(i, j)$, denoted by $S_{i,j}$, is given by

$$S_{i,j} = S_i \cup S_j.$$

Surname (dis)similarity between regions can be quantified by using different measures. Isonymy between regions $i$ and $j$ in $R$ (Zei, 1983) is defined as

$$I_{i,j} = \sum_{q \in S_i} \sum_{l \in S_j} p_{qi} p_{lj} = \sum_{q \in S_{ij}} p_{qi} p_{qj}, \text{ with } i, j \in R. \tag{2}$$

By convenience, we assume that $I_{i,i} = I_i$, for all $i \in R$.

However, other isonymic distance measures between a pair of locations $(i, j)$ can be derived from Equation (2). For instance, the Lasker distance (Lasker, 1977), namely $L_{i,j}$, is given by

$$L_{i,j} = -\log(I_{i,j}), \text{ with } i, j \in R. \tag{3}$$

Lasker distances measure similarities between regions, where large distances indicate less similarity in surname composition. There exist other measures apart from Lasker distance to quantify surname similarity in literature as, for instance, Nei distance.

## 3. ISONYMY BETWEEN GROUPS OF REGIONS: A COOPERATIVE GAME THEORY APPROACH

The main purpose of this work consists of ranking the Galician councils, taking a groundbreaking perspective from cooperative game theory. More specifically, the criterion chosen is based on their marginal capability to modify the overall isonymy. However, the most usual notions involve only two regions and, for this reason, a generalization of isonymy is required for their usage in a more general scenario of cooperation.

### *3.1. An extension of isonymy between groups*

We firstly generalize the definition of isonymy between in Equation (2) and Lasker distance in Equation (3) for a more general setting. Take again $R = \{1, \ldots, r\}$ the set of the involved geographical regions. For a subset $T$ of regions of $R$, if $S_j$ is the collection of surnames of region $j$, with $j \in R$, the joint collection of all the surnames is naturally given by

$$S_T = \bigcup_{j \in T} S_j.$$

The usage of isonymy to measure of population similarities of groups with more than 2 regions can be naturally justified. First, an extension of the concepts of isonymy between and Lasker

distance to a more general setting is necessary. If we take a subset of regions $T \subseteq R$, the isonymy between them can be innovatively prescribed by

$$I_T = \sum_{l \in S_T} \Big( \prod_{j \in T} p_{lj} \Big), \tag{4}$$

as an extension of Equation (2). Directly, different measures of the isonymic distance between a set of regions may be determined. For instance, the Lasker distance for $T$, namely $L_T$, is given by

$$L_T = -\log(I_T). \tag{5}$$

*3.2. Isonymy under cooperation*

Now, we will make use of cooperative game theory and its main results in order to rank regions. Formally, a *transferable utility cooperative game* (or TU-game) is given by $(R, v)$, with $R = \{1, \ldots, r\}$ the set of players and $v : 2^R \longrightarrow \mathbb{R}$ a map that assigns to each coalition $T \subseteq R$ a real number $v(T)$ that represents the worth of the cooperation of the members of $T$, with $v(\emptyset) = 0$. $G^R$ denotes the class of TU-games with set of players $R$.

Directly, the problem of cooperation of the regions can be modelled by TU-games. Take $T \subseteq R$ a group of regions of $R$. Thus, the isonymy games $(R, v^I)$ and $(R, v^L)$ can be alternatively defined for every $T \subseteq R$, when using Equation (4), by

$$v^I(T) \quad = \quad I_T$$

and, when using Equation (5), by

$$v^L(T) \quad = \quad L_T.$$

Note that a comprehensive analysis of the theoretical properties that these classes of TU-games satisfy can be done. However, they have not included here because it is out of the scope of the paper. If it was of interest, this analysis can be provided on request.

The establishment of rankings by using TU-games is an idea already considered in other settings. Mainly, specific solutions of TU-games are used for this purpose. It is remarkable the case of the Shapley value for a general TU-game $(R, v)$ (Shapley, 1953), whose definition is based on the average of those contributions of a player to the set of coalitions that do not contain it. That is, fixed $i \in R$ and $(R, v) \in G^R$, player $i$'s *marginal contribution* to coalition $T \subseteq R \setminus \{i\}$ is given by

$$v(T \cup \{i\}) - v(T). \tag{6}$$

Formally, the *Shapley value* can be expressed in terms of permutations. $\Pi(R)$ is the set of permutations of $R$. For each $\sigma \in \Pi(R)$, the set of predecessors of player $i \in R$ according to $\sigma$ is denoted by $P_i^\sigma$ and defined as $P_i^\sigma = \{j \in R : \sigma(j) < \sigma(i)\}$. Thus, the Shapley value can be written as

$$Sh_i(R, v) = \frac{1}{|\Pi(R)|} \sum_{\sigma \in \Pi(R)} (v(P_i^\sigma \cup \{i\}) - v(P_i^\sigma)), \text{ for every } i \in R \text{ and every } (R, v) \in G^R.$$

The main drawback concerning the Shapley value is computational, since its complexity increases exponentially with the number of players. There are many papers dealing with this issue from several perspectives. For instance, Castro et al. (2009) introduced a polynomial estimation procedure of the Shapley value based on sampling. In this paper, we adapt it for its application in an onomastic setting.

## 4. APPLICATION: THE CASE OF THE SURNAMES IN GALICIA

The data of the surnames of Galicia were extracted from the register of inhabitants, for the year 2011, provided by the Galician Institute of Statistics (IGE, http://www.ige.eu/). The analysed data corresponds to 20,754 different surnames for the 2,430,512 inhabitants of the 315 Galician councils. Thus, we do $R = \{1, \ldots, 315\}$.

Figure 1 depicts the ranking of the Galician councils according to their capabilities of modifying the overall isonymy in the case of a merger, after apply generalisation of isonymy between or Lasker distance. More specifically, it shows the coloured maps of Galicia resulting from the Shapley value estimation for the two isonymy games under consideration. In particular, light colours in councils represent shorter components of the estimated Shapley value, and darker colours represent the largest components.

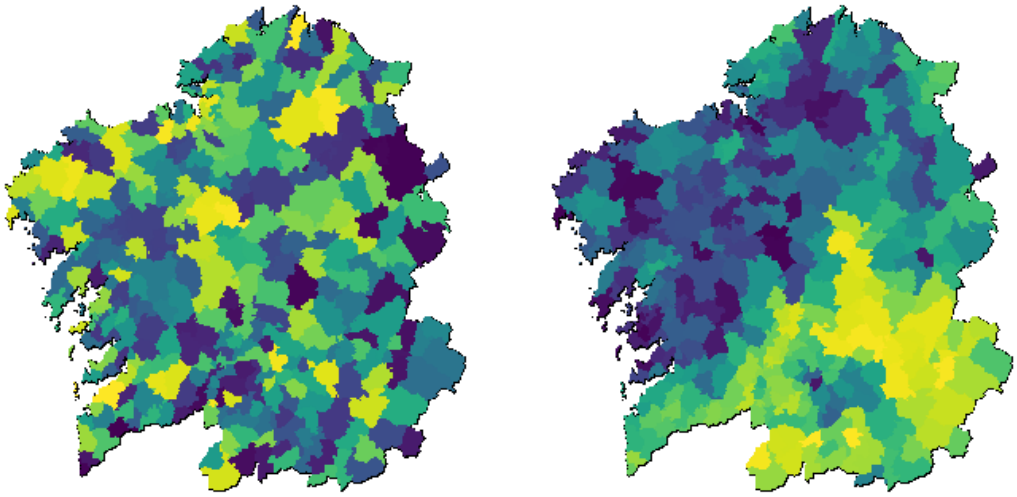


Figure 1: Results of the estimated Shapley value for $(R, v^I)$ (left) and for $(R, v^L)$ (right).

Isonymy is related to ecology, as applied to species, and thus to diversity. High isonymy values correspond to low diversity and vice versa, low isonymy values correspond to high diversity. From the obtined results when $(R, v^I)$, no clear pattern of town councils is observed in terms of their influence on the overall isonymy. However, we check that in the southern area of the province of Lugo and the province of Ourense obtain the shortest components of the estimated Shapley value for the case of the Lasker distance under this cooperative approach. That is, all these councils decrease on overage the overall isonymy.

It is noteworthy that similar conclusions can be extracted from Ginzo-Villamayor et al. (2013), which indicated scarcity (low diversity in ecological terms) of surnames in that area compared to the rest of Galicia. This area is characterized by the predominance of patronymic surnames. This class of surnames are those that ending in "-ez" and originate from a proper name. For example, González means "son of Gonzalo", Fernández, of Fernando, and Rodríguez, of Rodrigo, which is the most frequent surname in Galicia. It is remarkable the fact of that only nineteen of the most common surnames in Galicia end in "-ez" in spite of they represent more than 50% of the surnames in the Galician population. That is, there are not many surnames, but many people have them. Figure 2 shows results of Galician Surname Maps[1] for surnames corresponding to the "-ez" pattern. In Galicia, patronymic surnames are mainly found in those councils of the south of the province of Lugo and those of the province of Ourense. Clearly, there exists a correspondence between this map and the one obtained by colouring the Galician councils according to the decreasing order of the components of the estimated Shapley value for the isonymy game when considering the Lasker distance (see Figure 1, right).

[1]Galician Surname Maps (Cartografía dos Apelidos de Galicia, GSM) is a project of the Instituto da Lingua Galega of the Universidade de Santiago de Compostela to provide a research tool for the study of the geographical distribution of surnames in Galicia. This employs a geographical information system combining statistical data with a spatial analysis.

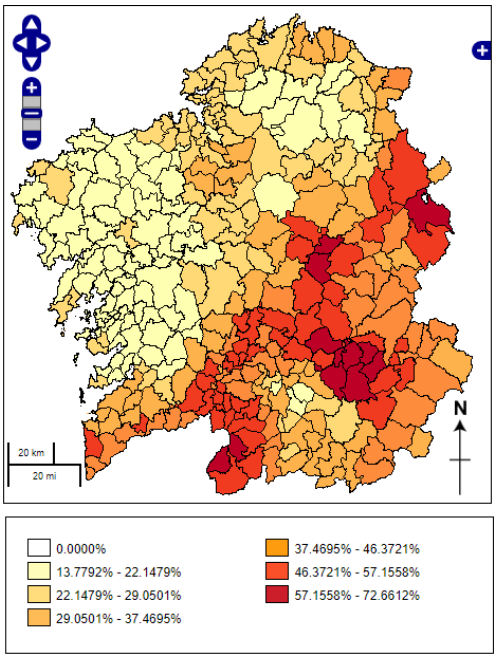



Figure 2: Results of Galician Surname Maps for surnames corresponding to the "-ez" pattern.

## REFERENCES

Boattini, A., Lisa, A., Fiorani, O., Zei, G., Pettener, D. and Manni, F. (2012). General method to unravel ancient population structures through surnames, final validation on Italian data. Human Biology, 84, 235-270.

Boattini, A., Pedrosi, M.E., Luiselli, D. and Pettener, D. (2010). Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). Annals of Human Biology, 37, 604-609.

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

Cheshire, J.A., Longley, P.A. and Singleton, A.D. (2010). The surname regions of Great Britain. Journal of Maps, 6, 401-409.

Ginzo-Villamayor, M.J., Crujeiras, R. M. and Sousa Fernández, X. (2013). Surname patterns in Galicia. Libro de Actas do XI Congreso Galego de Estatística e Investigación de Operacións. A Coruña (España).

Lasker, G. W. (1968). The occurrence of identical (isonymous) surnames in various relationships in pedigrees: a preliminary analysis of the relation of surname combinations to inbreeding. American Journal of Human Genetics, 20(3), 250.

Lasker, G. W. (1977). A coefficient of relationship by isonymy: A method for estimating the genetic relationship between populations. Human Biology, 49, 489-493.

Mikerezi, I., Shina, E., Scapoli, C., Barbujani, G., Mamolini, E., Sandri, M., Carrieri, A., Rodríguez-Larralde, A. and Barrai, I. (2013). Surnames in Albania: a study of the population of Albania through isonymy. Annals of Human Genetics, 77, 232-243.

Novotny, J. and Cheshire, J. A. (2012). The surname space of the Czech Republic: Examining population structure by network analysis of Spatial co-occurrence of surnames. PloS One. 7(10): e48568. doi:10.1371/journal.pone.0048568

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

Zei, G., Guglielmino, C.R., Siri, E., Moroni, A. and Cavalli-Sforza, L. L. (1983). Surnames as neutral alleles: Observations in Sardinia. Human Biology, 55(2), 357-365.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# BIVARIATE COPULA GENERALIZED ADDITIVE MODELS, FOR LOCATION, SCALE, AND SHAPE (CGAMLSS). APPLICATION TO PERINATAL MENTAL HEALTH (RISEUP-PPD-COVID-19 STUDY)

Carla Díaz-Louzao[1,2], Ana Mesquita[3], Raquel Costa[4,5,6], Emma Motrico[7], Francisco Gude[8] and Carmen Cadarso-Suárez[2]

[1]Department of Psychiatry, Radiology, Public Health, Nursing, and Medicine. University of Santiago de Compostela, Spain

[2]Department of Statistics, Mathematical Analisys, and Optimization, Group of Biostatistics and Biomedical Data Science. University of Santiago de Compostela, Spain

[3]School of Psychology. University of Minho, Portugal

[4]EPIUnit, Instituto de Saúde Pública. Universidade do Porto, Portugal

[5]Laboratório para a Investigação Integrativa e Translacional em Saúde Populacional (ITR), Porto, Portugal

[6]Human-Environment Interaction Laboratory. Universidade Lusófona do Porto, Portugal

[7]Psychology Department, Universidad Loyola Andalucia, Spain

[8]Department of Epidemiology, University Clinical Hospital of Santiago, Research Group on Epidemiology of Common Diseases, Santiago de Compostela Health Research Institute (IDIS), Santiago de Compostela, Spain.

## ABSTRACT

COVID-19 is a new pandemic, declared a public health emergency by the World Health Organization (WHO). According to the latest report of 28 September 2021, the cumulative number of confirmed cases reported globally is now over 231 million and the cumulative number of deaths is more than 4.7 million (WHO, 2021). This context is believed to have negative consequences for pregnant and postpartum women. The scarce evidence published to date suggests that perinatal mental health has deteriorated since the COVID-19 outbreak. Within the "Research Innovation and Sustainable Pan-European Network in Peripartum Depression Disorder - Riseup-PPD" (Cost Action 18138), an international prospective cohort study is underway to assess the impact of COVID-19 in perinatal mental health, carried out by the "Perinatal Mental Health and COVID-19 Pandemic" task force (Motrico et al., 2021).

The data come from online questionnaires to women in the perinatal period (pregnant or with a child of 6 months or less) with residence in 10 European countries (Albania, Bulgaria, Cyprus, Greece, Israel, Malta, Portugal, Spain, Turkey, and the United Kingdom), Brazil and Chile. These data collect demographic information, information on the participants' COVID-19 experience, and surveys to determine the level of anxiety and depressive. In addition, macro variables related to the epidemiological situation and public health measures regarding COVID-19, that are available for each participant's country in the Oxford COVID-19 Government Response Tracker (OxCGRT), were also used.

The main objective of this study is to determine possible factors that could be affecting perinatal mental health (anxiety and depression). As these two outcomes are known to be highly correlated, a Bivariate Copula Additive Model for Location, Scale, and Shape (CGAMLSS; Marra and Radice, 2017) is fitted. The advantages of this model are numerous. On the one hand, it does not restrict the response distributions to the exponential family. On the other hand, both the parameters of the marginals and the parameters of the copula itself can be made dependent on explanatory variables through flexible functions, including random effects (which, in this case, need to be considered as the data are aggregated by country). As a final remark, it should be noted that this technique allows the correlation between the responses to depend on covariates of interest.

**Keywords**: COVID-19, Depression, Anxiety, Bivariate regression, CGAMLSS.

## AKNOWLEDGEMENTS

The project is part of the COST Action Riseup-PPD CA 18138 and was supported by COST under COST Action Riseup-PPD CA18138. Carla Díaz-Louzao is supported by the Ministry of Economy and Competitiveness (Spain), and by the European Regional Development Fund (ERDF) under the project MTM2017-83513-R, and also under the project ED431C 2020/20, financed by the Competitive Research Unit Consolidation 2020 Programme of the Galician Regional Authority (Xunta de Galicia). Ana Mesquita is supported from the Portuguese Foundation for Science and Technology (FCT) and from EU through the European Social Fund and from the Human Potential Operational Program - IF/00750/2015. Raquel Costa was supported by the FSE and FCT under the Post-Doctoral Grant SFRH/BPD/117597/2016.

## REFERENCES

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. Computational Statistics & Data Analysis, 112, 99–113.

Motrico, E., Bina, R., Domínguez-Salas, S., Mateus, V., Contreras-García, Y., Carrasco-Portiño, M., Ajaz, E., Apter, G., Christoforou, A., Dikmen-Yildiz, P., et al. (2021). Impact of the Covid-19 pandemic on perinatal mental health (Riseup-PPD-COVID-19): protocol for an international prospective cohort study. BMC Public Health, 21(1), 1–11.

WHO (2021). Weekly epidemiological update on COVID-19 - 28 September 2021. Accessed on 29th September, 2021. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—28-september-2021.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# INCLUDING COVARIATES IN ROC CURVE ANALYSES

Arís Fanjul Hevia[1], Wenceslao González Manteiga[2] and Juan Carlos Pardo Fernández[3]

[1]Departamento de Estadística e Investigación Operativa y Didáctica de la Matemática, Universidad de Oviedo

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

[3]Departamento de Estatística e Investigación Operativa and Centro de Investigacións Biomédicas (CINBIO), Universidade de Vigo

## ABSTRACT

The Receiver Operating Characteristic (ROC) curve is a statistical tool that is used to asses the discriminatory capability of a classification problem such as the diagnostic method of a certain disease. Along with the diagnostic variables that are used for its construction, it is usual to have some other covariates that give extra information about the disease. These covariates could have an effect on the diagnostic variables or in their corresponding ROC curves and, thus, they should be incorporated into the analysis. Using the covariate-adjusted, the covariate-specific or the pooled ROC curves, we discuss how to decide if we can exclude the covariates from our study or not, and the implications this may have in further analyses of the ROC curve.

**Keywords**: comparison, covariates, hypothesis testing, ROC curves.

## 1. INTRODUCTION

Any diagnostic method is based on a classification problem in which the aim is to discriminate between two populations, usually identified as the healthy population and the diseased population, in such a way that the number of subjects that are misclassified is minimized. The Receiver Operating Characteristic (ROC) curves are commonly used in this context for studying the behaviour of the classification variables. See, for example, the monograph of Pepe (2003) as an introduction to the topic.

The most common way of representing this curve is by the formula:

$$ROC(p) = 1 - F(G^{-1}(1-p)), \quad 0 \leq p \leq 1,$$

where $F$ and $G$ are the cumulative distribution functions of the diagnostic variable in the diseased and in the healthy population, respectively.

In a practical situation it is usual to have some other covariates, either continuous (such as blood pressure, age or body mass index of the patients) or discrete (such as gender, medical history, hospital where the treatment is given,...). This situation raises the question of whether this extra information should or not be included in the ROC curve analysis. In this work we discuss how to answer this question and the implications it may have for other methodologies that involve ROC curves, providing an example with real data to illustrate the problem.

## 2. ASSESSING THE COVARIATE EFFECT

Three different situations can arise when dealing with ROC curves with covariates: in the first one the performance of the ROC curve changes with the value of the covariates (and with it, its discriminatory capability); in the second, the covariates affect the distribution of the diagnostic markers, but not their discriminatory capability; in the last one, the covariates do not affect the ROC curve in any way.

Apart from the strategy of ignoring any covariate information that we may have (which means using the pooled data to conform the *pooled ROC curve*), there are two ways of modelling the effect of a covariate in an ROC curve: using the *covariate-specific* or *conditional ROC curve* or using the *covariate-adjusted ROC curve*.

For a fixed value of the covariate $x \in R_X$, where $R_X$ is the support of the continuous covariate $X$, the conditional ROC curve is defined as

$$ROC^x(p) = 1 - F(G^{-1}(1-p|x)|x),\ p \in (0,1),$$

where $F(\cdot|x)$ and $G(\cdot|x)$ are the cumulative distribution functions of the diagnostic variable in the diseased and healthy population, respectively, conditioned to the value $x$. Note that its structure is very similar to the previously defined pooled ROC curve, except that the distribution functions are now conditioned to the value of the covariate $x$.

The *covariate-adjusted ROC curve* (AROC curve) (first introduced by Janes and Pepe, 2009), which can be viewed as a weighted average of conditional ROC curves, is defined as

$$AROC(p) = 1 - F(G^{-1}(1-p|X^F)),\ p \in (0,1).$$

There are several documents in the literature dedicated to the estimation of these curves (a review of them can be found in Pardo-Fernández et al. 2014). Studying the relationship that can be established between these three curves, we can decide which one should be used in each scenario. Deciding the situation we have at hand is a two-step problem.

The first step should be to test whether the conditional ROC curve is constant for each value of the covariate $x \in R_X$, meaning $ROC^z(p) = ROC^x(p)\ \forall x \in R_X$, with $p \in (0,1)$, for a certain fixed value $z \in R_X$. In this case, this $ROC^z$ would coincide with the covariate-adjusted ROC curve. Thus, we would be interested in testing:

$$H_0^1 : ROC^x(p) = AROC(p),\ p \in (0,1)\ \forall x \in R_X, \tag{1}$$

versus the general alternative $H_1^1 : H_0^1$ is not true.

If this null hypothesis were to be rejected, we would be in a scenario where the discriminatory capability of the diagnostic marker is affected by the covariate. In this case we should use the conditional ROC curve for further analysis.

Otherwise, one could think about eliminating the covariate from the analysis, but accepting that $ROC^x = AROC$ for any $x$ does not necessarily mean that the pooled ROC curve is going to coincide with the AROC curve. Thus, one should make this test:

$$H_0^2 : AROC(p) = ROC(p),\ p \in (0,1), \tag{2}$$

versus the general alternative $H_1^2 : H_0^2$ is not true. If this hypothesis were rejected, the AROC curve should be considered.

Only if both of the above mentioned hypotheses, $H_0^1$ and $H_0^2$, hold can we consider removing the covariates from the analysis (using, thus, the pooled ROC curve).

There are not many references regarding these problems in the literature. Rodríguez-Álvarez et al.(2011) propose a test for dealing with the first test (1) in a case with a unidimensional covariate and Rodríguez-Álvarez et al. (2018) propose an inferential procedure for testing the effect of covariates over the conditional ROC curve employing generalized additive models, using two different bootstrap-based tests to check the possible effect of the continuous covariates on the ROC curve and the presence of factor-by-curve interaction terms. In this work we propose a new methodology to test the second hypothesis (2), focusing in the case with only one covariate.

## 3. FURTHER GOALS

Apart of the use of the ROC curve for the evaluation of diagnostic variables, the analysis of these sort of curves can be used for further goals. For example, when there are two or more diagnostic methods for a certain disease, their performance can be compared through the comparison of their corresponding ROC curves (Fanjul-Hevia and González-Manteiga, 2018).

That said, when information about covariates is available it should be incorporated in the comparison. This can be done through the comparison of the conditional ROC curves (Fanjul-Hevia et al. 2020 and Fanjul-Hevia et al., 2021), but depending on the result of the significance of the covariate effect on the ROC curves this could also mean to compare pooled or covariate-adjusted ROC curves.

## ACKNOWLEDGEMENTS

A. Fanjul-Hevia and W. González-Manteiga acknowledge the support from the Spanish Ministerio de Economía, Industria y Competitividad, through grant number and MTM2016-76969-P, which includes support from the European Regional Development Fund (ERDF). J.C. Pardo-Fernández acknowledges financial support from grant PID2020-118101GB-I00, funded by the Spanish Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación and the ERDF.

## REFERENCES

Fanjul-Hevia A. and González-Manteiga W. (2018). A comparative study of methods for testing the equality of two or more ROC curves. *Computational Statistics*, 33, 357–377.

Fanjul-Hevia A., González-Manteiga W. and Pardo-Fernández J.C. (2020). A non parametric test for comparing conditional ROC curves. *Computational Statistics & Data Analysis*, 157: 107146.

Fanjul-Hevia A., Pardo-Fernández J.C., Van Keilegom I. and González-Manteiga W. (2021). A test for comparing conditional ROC curves with multidimensional covariates. Manuscript submitted for publication.

Janes H. and Pepe M.S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* 96, 371–382.

Pardo-Fernández J.C., Rodríguez-Álvarez M.X. and Van Keilegom I. (2014). A review on ROC curves in the presence of covariates. REVSTAT- *Statistical Journal*, 12, 21–41.

Pepe M.S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press.

Rodríguez-Álvarez M.X., Roca-Pardiñas J. and Cadarso-Suárez C. (2011). ROC curve and covariates: extending induced methodology to the non-parametric framework.*Statistics and Computing*, 21, 483–499.

Rodríguez-Álvarez M. X., Roca-Pardiñas J., Cadarso-Suárez C. and Tahoces P. G. (2018). Bootstrap-based procedures for inference in nonparametric receiver operating characteristic curve regression analysis. *Statistical Methods in Medical Research*, 27, 740–764.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Predicción Cooperativa en el contexto de Covid-19

Manuel Antonio Novo Pérez[1], Víctor González Carro[1], Carlos Fernández Lozano[1], José Antonio Vilar Fernández[1], Luis Ángel García Escudero[2], Pablo Montero Manso[3] y Rubén Fernández Casal[1].

[1]CITIC
[2]Universidad de Valladolid
[3]University of Sidney

## RESUMEN

Como parte de la iniciativa "Acción matemática contra el Coronavirus" promovida por el Comité Español de Matemáticas (CEMat), se desarrolló una aplicación web basada en `R` para monitorizar y predecir el comportamiento a corto plazo de ciertas variables relevantes en la transmisión de la Covid-19. Para cada Comunidad Autónoma, se combinaron predicciones de diferentes modelos y técnicas proporcionados por grupos de investigación independientes para generar predicciones cooperativas, que han estado disponibles en la web diariamente junto con las series oficiales de datos del Instituto de Salud Carlos III (ISCIII). Puesto que la combinación de predicciones puede mejorar la precisión de la predicción, especialmente en contextos de incertidumbre, el objetivo de este proyecto era usar las predicciones de la comunidad matemática española para obtener predicciones más precisas y estables y, finalmente, reportar conclusiones a las autoridades. Este trabajo proporciona un análisis del desarrollo y los resultados, motivando el uso de predicciones combinadas e incluyendo información de las principales etapas del proceso.

**Palabras y frases clave**: Covid-19, Precisión de las predicciones, Combinación de predicciones, Modelos de predicción, Pesos óptimos

## 1. INTRODUCCIÓN

A finales de 2019, se empezaron a conocer en China los primeros casos de la enfermedad infecciosa coronavirus 2019 (Covid-19). Su origen es un virus desconocido de la familia de coronavirus, el coronavirus 2 del síndrome respiratorio agudo severo (SARSCoV-2), y su preocupante nivel de propagación llevó a la Organización Mundial de la Salud (OMS) a declarar oficialmente la enfermedad como pandemia el 11 de marzo de 2020. A finales de julio de 2021, más de un año y medio después, la información oficial proporcionada por el Ministerio de Sanidad muestra que el número de casos confirmados en todo el mundo desde el inicio de la pandemia llega casi hasta los 200 millones de casos, con más de 4 millones de fallecidos.

Desde entonces, los esfuerzos de los gobiernos y la comunidad científica se han centrado en el estudio y detención de la expansión del virus. En particular, desde el Comité Español de Matemáticas (CEMat) surgió una iniciativa denominada *Acción matemática contra el coronavirus*, cuyo objetivo era poner a disposición de las autoridades la capacidad de análisis y modelización matemática para comprender el comportamiento de la Covid-19. Alineado con esta iniciativa de CEMat, se creó por parte del mismo organismo el proyecto de *Predicción Cooperativa*, que tenía como objetivo la creación de una web que recogiese tanto los datos publicados por los organismos oficiales como meta-predicciones sobre diferentes variables de interés relacionadas con el coronavirus.

El meta-predictor o predictor cooperativo se construye a partir de combinaciones de predicciones basadas en diferentes modelos o algoritmos. Con este fin, el 1 de abril de 2020 se hizo un

llamamiento a todas las personas en disposición de elaborar modelos para predecir la evolución de la enfermedad, solicitando su colaboración en la construcción de meta-predicciones mediante el envío diario de sus propias predicciones. En el contexto de abril de 2020, habida cuenta de la situación sanitaria en la que se encontraba el país, resultaba de gran utilidad disponer de predicciones acerca de variables relevantes a corto y medio plazo. No obstante, al ser la Covid-19 una enfermedad nueva, el mecanismo generador de los datos era totalmente desconocido, de modo que los riesgos de que un determinado modelo esté sesgado y funcione para una determinada muestra y no para otras son elevados. Una buena forma de reducir este riesgo es combinar predicciones de diferentes modelos, lo cual fue la principal motivación de esta iniciativa.

Las predicciones se realizaban diariamente y se hacía una predicción para cada Comunidad Autónoma (incluidas Ceuta, Melilla y el total de España), variable y horizonte. Las variables consideradas fueron número de casos confirmados, número de hospitalizaciones, número de ingresos en UCI, número de fallecimientos y número de nuevos casos. Por su parte, los horizontes a considerar iban de 1 a 7 días. Cabe destacar que para las cuatro primeras variables se predecía la cifra acumulada y que los predictores individuales podían mandar las predicciones para las variables, horizontes y territorios que considerasen oportunos, no era obligatorio predecir todos los escenarios.

Actualmente, *Predicción Cooperativa* sigue realizando predicciones y además se están analizando con detalle los resultados obtenidos en la primera etapa para sacar conclusiones. En este sentido, la metodología actual ha cambiado ligeramente respecto a la etapa del año 2020, siendo las principales diferencias las siguientes:

- Ahora las predicciones se hacen un día a la semana y se predice a 14 horizontes en lugar de 7.
- Ya no se predice la variable "Nuevos casos".
- Los datos proporcionados por el Instituto de Salud Carlos III son tremendamente inestables, ya que las series históricas se modifican todos los días. Esto complica mucho la evaluación, y se ha optado por que los datos queden inmóviles tras un tiempo (8 días, que es el tiempo medio que tardan en estabilizarse las series). Así, por ejemplo, si se hace una predicción para el día 15 de julio, no se evaluará con el dato publicado el día 16, como cabría esperar, sino con el valor observado para el día 15 en el archivo publicado el día 23.

## 2. PREDICTORES COOPERATIVOS

De aquí en adelante, $\hat{y}_{i,t+h|t}$ denota la predicción para el horizonte $h$ realizada en el día $t$ por el $i$-ésimo predictor y $npre$ es el número total de predictores para cada escenario, considerando como escenario cada combinación de variable, territorio y horizonte. Los predictores cooperativos utilizados fueron los siguientes:

- **CP01: Media muestral.** Cada predictor recibe el mismo peso:

$$\hat{y}_{CP01,t+h|t} = \frac{1}{npre}\sum_{i=1}^{npre}\hat{y}_{i,t+h|t}.$$

- **CP02: Mediana muestral.** Una alternativa más robusta a la media muestral:

$$\hat{y}_{CP02,t+h|t} = \begin{cases} \hat{y}_{\left(\frac{npre+1}{2}\right),t+h|t} & \text{si } npre \text{ impar} \\ \frac{1}{2}\left(\hat{y}_{\left(\frac{npre}{2}\right),t+h|t} + \hat{y}_{\left(\frac{npre}{2}+1\right),t+h|t}\right) & \text{si } npre \text{ par} \end{cases}$$

donde $\hat{y}_{(i),t+h|t}$ es la observación que ocupa el $i$-ésimo lugar de la muestra ordenada ascendentemente.

- **CP03: Media recortada.** Otra alternativa robusta que consiste en calcular la media tras eliminar las $100 \cdot \lambda\%$ observaciones más extremas, con $0 < \lambda < 1$. Si $K = \lambda \cdot npre$:

$$\hat{y}_{CP03,t+h|t} = \frac{1}{npre - 2K} \sum_{i=K+1}^{npre-K} \hat{y}_{(i),t+h|t}.$$

  En este caso se tomó $\lambda = 0.2$.

- **CP04: Media Windsorizada.** Consiste en sustituir los $100 \cdot \lambda\%$ valores más extremos por el valor más extremo de los valores restantes. Si, como antes $K = \lambda \cdot npre$:

$$\hat{y}_{CP04,t+h|t} = \frac{1}{npre} \left( K \left( \hat{y}_{(K+1),t+h|t} + \hat{y}_{(npre-K+1),t+h|t} \right) + \sum_{i=K+1}^{npre-K} \hat{y}_{(i),t+h|t} \right).$$

  Nuevamente, se consideró $\lambda = 0.2$.

- **CP05: Bates/Grander (mod).** Predicciones combinadas con pesos normalizados. Estos pesos son más elevados cuando el predictor es más preciso (es decir, menor error en los días previos). Seguimos el procedimiento de Bates y Granger pero con diferentes pesos. Más específicamente:

$$\hat{y}_{CP05,t+h|t} = \sum_{i=1}^{nphis} \omega_{i,t+h|t}, \hat{y}_{i,t+h|t}, \text{ being } \omega_{i,t+h|t} = \frac{1/\hat{\varepsilon}_{i,t+h|t}}{\sum_{i=1}^{nphis} 1/\hat{\varepsilon}_{i,t+h|t}} \quad (1)$$

  donde $nphis$ es el número de predictores que mandaron predicciones en los días previos a $t-h, (\hat{y}_{t-1|t-1-h}, ..., \hat{y}_{t-n_i|t-n_i-h})$ y $\hat{\varepsilon}_{i,t+h|t}$ el error medio del $i$-ésimo predictor en los días previos.

  Para un día $t$ y un escenario (variable, territorio, horizonte), $\hat{\varepsilon}_{i,t+h|t}$ se calcula como sigue.

  1. Sea $ndhis = \max_{1 \le i \le nphis} n_i$, el mayor número de días previos a $t-h$ con alguna predicción a horizonte $h$. Sea $\mathcal{M}_h$ la matriz de dimensión $nhdis \times nphis$ con las predicciones previas, cuyo elemento $(j, i)$ se define como $\hat{y}_{i,t-j|t-j-h}$. La primera fila y al menos una columna de $\mathcal{M}$ tendrán todas las predicciones, pero en la mayoría de los casos $\mathcal{M}$ no estará completa.

  2. Sean $\boldsymbol{y} = (y_{t-1}, ..., y_{t-ndhis})^T$ el vector de valores reales de la serie los $ndhis$ días anteriores a $t$ $\mathbf{1} = (1, ...1)^T$ el vector de $nphis$ unos, entonces la matriz $\mathcal{E}_h = |\mathbf{1}\boldsymbol{y}^{\boldsymbol{T}} - \mathcal{M}_h| = (e_{ji,h})$, con $e_{ji,h} = |y_{t-j} - \hat{y}_{i,t-j|t-j-h}|$ contiene los errores absolutos de los predictores en el histórico. Los valores que faltan en $\mathcal{E}_h$ se imputan con $\max_{1 \le i \le ndhis} (nphis)^{-1} \{e_{ji,h}\}$.

  3. El error medio $\hat{\varepsilon}_{i,t+h|t}$ del $i$-ésimo predictor se obtiene a partir de las columnas de $\mathcal{E}_h$ usando alguno de los siguientes criterios:

| | |
|---|---|
| MAE (Mean Absolute Error) | $\hat{\varepsilon}_{i,t+h|t} = \frac{1}{nhdis} \sum_{k=1}^{nhdis} e_{ji,h}$ |
| RMSE (Root Mean Squared Error) | $\hat{\varepsilon}_{i,t+h|t} = \left( \frac{1}{nhdis} \sum_{k=1}^{nhdis} e_{ki}^2 \right)^{1/2}$ |
| MAPE (Mean Absolute Percentage Error) | $\hat{\varepsilon}_{i,t+h|t} = \frac{1}{nhdis} \sum_{k=1}^{nhdis} (e_{ji,h}/y_{t-k})$ |

  Nótese que CP05 ignora la estructura de covarianza entre los errores. Se ha optado por esta vía para evitar un incremento en la varianza de las estimaciones de los pesos debida a predicciones fuertemente correladas.

  Además, intentando minorar el efecto de los cambios que se producían en los registros oficiales, se implementó una modificación de CP05 dando más peso a los errores correspondientes a los registros más recientes.

- **CP06: Lowess.** Predicciones obtenidas mediante un suavizado robusto por regresión polinómico local de todas las predicciones individuales a lo largo de los siete horizontes de predicción. Se emplea el algoritmo Lowess (Locally weighted scatterplot smoothing), que realiza iterativamente un ajuste local lineal de las predicciones de modo que en cada iteración se introducen pesos de robustez penalizando a aquellas predicciones que en el ajuste previo generaron residuos elevados.

- **CP07: Loess + Bates/Granger (mod).** Como en el predictor CP06, se realiza un ajuste polinómico local ponderado de las predicciones pero empleando ahora el algoritmo Loess e introduciendo los mismos pesos que se obtuvieron en CP05 para los diferentes predictores individuales.

### 3. Conclusiones

El objetivo inicial del proyecto era desarrollar una aplicación web que mostrase los datos y las predicciones combinadas relacionadas con el coronavirus. Eso se consiguió en poco tiempo debido al esfuerzo desinteresado de muchas personas. Actualmente, la web sigue funcionando y contiene una mayor cantidad de información a través de la inclusión de los datos de vacunación o las cifras por edad y sexo.

Los resultados obtenidos muestran que, en este contexto, la combinación de predicciones es una buena idea. En algunos contextos puntuales, hay ciertos predictores individuales que destacan sobre el resto, pero son casi siempre los predictores cooperativos los que ocupan los primeros puestos. En particular, la mediana muestral (CP02) y la media recortada (CP03) obtienen generalmente los mejores resultados, sobre todo el primero de ellos.

Por último, las principales dificultades para llevar este proyecto a cabo están relacionadas con los datos. Esto pone de manifiesto la necesidad de disponer de unos datos consistentes, rigurosos y fiables.

## Referencias

[1] Aiolfi, M., and Favero, C. A. (2005). Model uncertainty, thick modeling and the predictability of stock returns. *Journal of Forecasting, 24(4), 233-254.*

[2] Aiolfi,M., and Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies (2006). *Journal of Econometrics, 135(1), 31-53.*

[3] Armstrong, J. S. (2001). Combining forecasts. *In: Principles of Forecasting: A Handbook for Researchers and Practitioners (Chap. 4), 417-439, Kluwer Academic Publishing, Dordrecht (Netherlands).*

[4] Bates, J. M., and Granger, C. W. J. (1969) The Combination of Forecasts. *Journal of the Operational Research Society, 20(4), 451-468.*

[5] Claeskens, G., Magnus, J. R., Vasnev, A. L. and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting, 32(3), 754-762.*

[6] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5(4), 559-583.*

[7] Clemen, R. T., and Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics, 4(1), 39-46.*

[8] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74(368), 829-836.*

[9] Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician, 35(1), 1-54.*

[10] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *In Statistical Models in S (Chap. 8). Eds. J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole, Pacific Grove, California, USA.*

[11] Cramer, Estee Y.; Lopez, Velma K.; Niemi, Jarad; George, Glover E.; Cegan, Jeffrey C.; Dettwiller, Ian D.; England, William P.; Farthing, Matthew W.; Hunter, Robert H.; Lafferty, Brandon; Linkov, Igor; Mayo, Michael L.; Parno, Matthew D.; Rowland, Michael A.; Trump, Benjamin D.; Wang, Lily; Gao, Lei; Gu, Zhiling; Kim, Myungjin; Wang, Yueying; Walker, Jo W.; Slayton, Rachel B.; Johansson, Michael; Biggerstaff, Matthew; and et al.(2021). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *Statistics Publications. 315.*

[12] Diebold, F. X., and Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting, 6(4), 503-508.*

[13] Elliott, G., and Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics, 122(1), 47-79.*

[14] Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics, 177(2), 357-373.*

[15] Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: can anything beat the simple average? *International Journal of Forecasting, 29(1), 108-121.*

[16] Graefe, A., Armstrong, J. S., Jones, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *Combining forecasts: An application to elections.*

[17] Granger, C. W. J., and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting, 3(2), 197-204.*

[18] Granger, C.W. J., and Jeon, Y. (2004). Thick modeling. *Economic Modelling, 21(2), 323-343.*

[19] Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics, 146(2), 342-350.*

[20] Hsiao, C., and Wan, S. K. (2014). Is there an optimal forecast combination?. *Journal of Econometrics, 178(2), 294-309.*

[21] Jose, V. R. R., and Winkler, R. L. (2008). Simple robust averages of forecasts: some empirical results. The M3-competition: Results, conclusions and implication. *International Journal of Forecasting, 24(1), 163-169.*

[22] Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association, 106(495), 1053-1066.*

[23] Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implication. *International Journal of Forecasting, 16(4), 451-476.*

[24] Marcellino, M. (2004). Forecast pooling for short time series of macroeconomic variables. *Oxford Bulletin of Economic and Statistics, 66(1), 91-112.*

[25] Newbold, P., and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A, 137(2), 131-165.*

[26] Smith, J., and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics, 71(3), 331-355.*

[27] Stock, J. H., and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23(6), 405-430.*

[28] Timmermann, A. (2006). Forecast combinations. *In: Handbook of Economic Forecasting (Chap. 4), 1, Elsevier/North-Holland, Amsterdam (Netherlands).*

[29] Vul, E., and Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science, 19(7), 645-647.*

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

[31] Robert L., Winkler, R. L., and Makridakis, S. (1983). The Combination of Forecasts. *Journal of the Royal Statistical Society. Series A, 146(2), 150- 157.*

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# CONTRASTE DE HIPÓTESE PARA O EFECTO DE COVARIABLES SOBRE REXIÓNS DE REFERENCIA BIVARIADAS

Lado-Baleato, Óscar[1], Roca-Pardiñas, Javier[2] e Cadarso-Suárez, Carmen[3]

[1,3] Departamento de Estatística, Análise Matemática, e Optimización, Grupo de Bioestatístca e Ciencia de Datos Biomédicos. Universidade de Santiago de Compostela.
[2] Departmento de Statistics and Operations Research (SiDOR). Universidade de Vigo.

## RESUMO

Os intervalos de referencia úsanse de forma rutineira na atención primaria, e hospitalaria, para determinar se o valor dun *test* diagnóstico continúo é "*normal*", ou se corresponde a un estado patolóxico. As rexións de referencia, constitúen a extensión desta regra diagnóstica ó contexto multivariado. Mentres que un intervalo de referencia está definido por dous puntos entre os cales se atopan a maioría dos pacientes sans, unha rexión de referencia definese por unha envoltura convexa que contén unha proporción elevada dos resultados multivariados dos mesmos. En enfermidades cuxo diagnóstico require a determinación de dous, ou máis, marcadores continúos correlacionados (p. ex., diabetes, hipotiroidismo) a aplicación dunha rexión de referencia ofrece maior sensibilidade, e especificidade, que o uso de varios intervalos univariados (Boyd, 2004). A pesar destas vantaxes, nos últimos 50 anos as rexións de referencia multivariadas aplicaronse en contadas ocasións na práctica clínica.

Coa finalidade de propoñer regras diagnósticas máis específicas, e precisas, é habitual considerar o efecto de certas características do paciente sobre a distribución do marcador na poboación san. Así mesmo, no contexto multivariado, a distribución conxunta de dúas variables correlacionadas pode estar condicionada polo valor de certas covariables. Nesta comunicación presentarase un contraste de hipótese para avaliar a significación estatística do efecto de covariables sobre unha rexión de referencia definida para unha variable bivariada continúa. O estatístico de contraste basease nunha distancia euclidea estandarizada polo perímetro da rexión, e a aproximación da súa distribución fíxose mediante un esquema de remostraxe. En estudos de simulación comprobamos que o contraste proposto respecta os niveis nominais en caso de hipótese nula (non efecto), e mostra unha curva de potencia satisfactoria.

Esta proposta complementa un modelo de regresión para a estimación de rexións de referencia de forma non-paramétrica, previamente publicado polos autores (Lado-Baleato et al., 2021). Para ilustrar a utilidade do contraste na práctica clínica avaliamos o efecto da idade e xénero sobre a distribución bivariada do peso e talla, dunha cohorte de nenos sans. O contraste proposto permitiu identificar que a rexión de referencia para o (peso – talla) depende dunha interacción entre a idade e xénero dos infantes. Ademáis, o uso dunha rexión de referencia, en comparación coas curvas percentís do índice de masa corporal, ofrece unha caracterización máis axeitada dos valores de "*normalidade"* para ambas medidas antropométricas.

**Palabras e frases chave:** rexións de referencia multivariadas, probas diagnósticas, regresión non-paramétrica.

## REFERENCIAS

Boyd J.C. (2004) Reference regions of two or more dimensions. Clinical Chemistry Laboratory Medicine, 42(7),739–746.

Lado-Baleato, Ó., Roca-Pardiñas, J., Cadarso-Suárez, C. e Gude, F. (2021) Modeling conditional reference regions: Application to glycemic markers. *Statistics in Medicine, doi:10.1002/sim.9163.*

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Analysis and prediction of indoor $CO_2$ levels with Functional Data Analysis for the prevention of SARS-CoV-2 infection

Víctor Teodoro[1], María Jesús Hernández[1], Carlos J. Escudero[1], Manuel Oviedo[1], Oscar Fontenla-Romero[1]

[1] Centro de Investigación en Tecnologías de la Información y la Comunicación (CITIC), Universidade da Coruña.

## ABSTRACT

The concentration of carbon dioxide ($CO_2$) in closed spaces is a good indicator of its ventilation rate, which is correlated with the concentration of aerosols, through which SARS-CoV-2 is spread. In this paper, we analyze the levels of $CO_2$ measured in university classrooms using functional data analysis techniques such as Functional Linear Regression, Functional ANOVA, among others, due to the fact that these variables have a continuous nature, and are observed through time. By doing this, we aim to predict dangerous levels of $CO_2$ that could imply an increase in the risk of airborne virus transmission or other cognitive problems. The aforementioned analysis is done over a dataset collected in the Universidade da Coruña during the three days of EBAU, which is a national test that Spanish students are required to take to enter university.

**Keywords**: $CO_2$ concentration level, SARS-CoV-2 prevention, Functional data analysis.

## 1. INTRODUCTION

The risk of SARS-CoV-2 infection and its prevention is a main concern in present time society. Many studies have been done on the virus airborne transmission and its connection to ventilation rate in indoor spaces to reduce the risk of infection. As shown in some studies like Peng and Jimenez, 2021, Batterman, 2017, Rudnick and Milton, 2003, indoor $CO_2$ concentration levels gives us a good indicator of this ventilation rate and therefore its study and control can give us a good chance to minimize indoor virus transmission. This is not only of interest for SARS-CoV-2 prevention, as it is known, Allen et al., 2016, that bad ventilation is linked to cognitive issues.

Throught this paper, we will try to make predictions of future indoor $CO_2$ concentration levels based in previously available information of $CO_2$ levels, occupation, wind speed and classroom volume. We will focus on predicting the mean $CO_2$ level in the next hour.

As the data utilized here consists of continuous variables recorded over time, it is of interest to treat them as functions of time, enabling us to make use of functional data analysis techniques.

This discussion will proceed as follows: on **Section 2** we will introduce a brief description of the statistical problem we are approaching and give some information about the data at our disposal. Next, in **Section 3** we will visualize the available data and make some comments on it, in form of an exploratory analysis. Finally, in **Section 4** we will provide some models to predict the mean concentration level of $CO_2$. For each proposed model, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) will be given. This error metrics are calculated with 10-fold cross validation. In addition, Akaike Information Criterion (AIC) and $R^2_{adj}$ are given, computed fitting the model with all the available data.

## 2. DESCRIPTION OF THE PROBLEM

As mentioned above, in this article we will construct models utilizing functions of time (functional data). Therefore, for a given hour of the day, $h \in 0, 1, 2, \ldots, 23$, we will have a function of time, $CO_{2h}(t),\ t \in [0, 60)$, that will give us the values the $CO_2$ concentration levels took during that hour. For this purpose, we will make use of stochastic processes of functional variables, and also stochastic processes taking values in $\mathbb{R}$.

The statistical problem we will approach is to give predictions of the mean level of $CO_2$ in the next hour, $\mathbb{E}[CO_{2h+1}]$, which is the mean function of a scalar stochastic process. And we will do so by using functional stochastic processes of $CO_2$ at the current hour, $CO_{2h}(t),\ t \in [0, 60)$ among other variables.

### 2.1. ON THE DATA

In this paper, we will be working with a dataset consisting of observations of the following variables and their measurement units: $CO_2$ ($ppm$), temperature (°C), atmospheric pressure (hPa), humidity (%), wind speed ($km/h$), volume of classroom ($m^3$), maximum ocuppation in a hour (number of people).

The $CO_2$ concentration has been measured with sensors during the three days of the EBAU in the Universidade da Coruña. A sensor was placed in each of the classrooms were the exams took place and the values of the variables were recorded with a sample period of a minute from 8 a.m. to 9. pm. In total, we have the measurements of 52 sensors placed in classrooms for 10 faculties distributed across 3 campuses.

The wind speed and the maximum ocuppation data are available for every hour. Together with the volume, this values were recorded for each individual classroom.

## 3. DATA VISUALIZATION

On the first day of the EBAU there was a presentation from 9:00 am to 10:00 am explaining the examination procedures. After that, exams took place with the schedule seen in Table 1. It is to mention that students did not take all these exams, they only took those corresponding to their studies and choice, so the examination classrooms ocuppation varied among exams.

| Exam | Date | From | To |
|---|---|---|---|
| History of Spain | 8/6/2021 | 10:00 | 11:30 |
| Spanish Language and Literature | 8/6/2021 | 12:00 | 13:30 |
| Mathematics Applied to the Social Sciences | 8/6/2021 | 15:30 | 17:00 |
| Arts Fundamentals | 8/6/2021 | 15:30 | 17:00 |
| Enterprise Economy | 8/6/2021 | 17:30 | 18:00 |
| Design | 8/6/2021 | 17:30 | 18:00 |
| Galician Language and Literature | 9/6/2021 | 9:00 | 10:30 |
| First Foreing Language | 9/6/2021 | 11:00 | 12:30 |
| Mathematics | 9/6/2021 | 13:00 | 14:30 |
| Latin | 9/6/2021 | 13:00 | 14:30 |
| Technical Drawing | 9/6/2021 | 16:00 | 17:30 |
| Scenical Arts | 9/6/2021 | 16:00 | 17:30 |
| Chemistry | 9/6/2021 | 18:00 | 19:30 |
| Greek | 9/6/2021 | 18:00 | 19:30 |
| Biology | 10/6/2021 | 9:00 | 10:30 |
| History of Art | 10/6/2021 | 9:00 | 10:30 |
| Physics | 10/6/2021 | 11:00 | 12:30 |
| Geography | 10/6/2021 | 11:00 | 12:30 |
| Audiovisual Culture | 10/6/2021 | 11:00 | 12:30 |
| Geology | 10/6/2021 | 13:00 | 14:30 |
| History of Philosophy | 10/6/2021 | 13:00 | 14:30 |

Table 1: Exam schedules

In Figure 1, the evolution of the $CO_2$ concentration levels recorded by one sensor in one of the exam classrooms on the first day of the EBAU can be seen. We can clearly see that, as time passes during an exam, the $CO_2$ concentration level increases, decreasing as people finish the exam and leave the classroom. We can also see that it would be beneficial to have predictions of the levels reached in each exam, as for example, on the first exam of this particular day a peak of slightly more than $900ppm$ is reached and under some circumstances this could contribute to airborne virus transmission. If we had a model to predict this level, professors could had been alerted to better ventilate this classroom and prevent it.

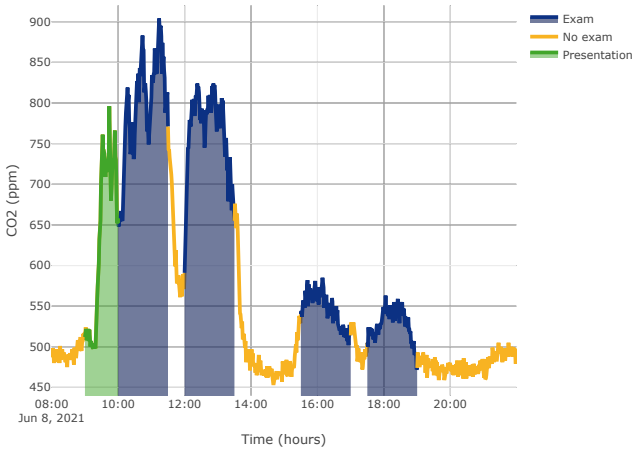



Figure 1: $CO_2$ levels of one of the sensors on Universidade da Coruña's Law School for the first day of EBAU

In Figure 2, we can observe the spaghetti plot for the visualization of all our $CO_2$ concentration functions for the three days of EBAU.

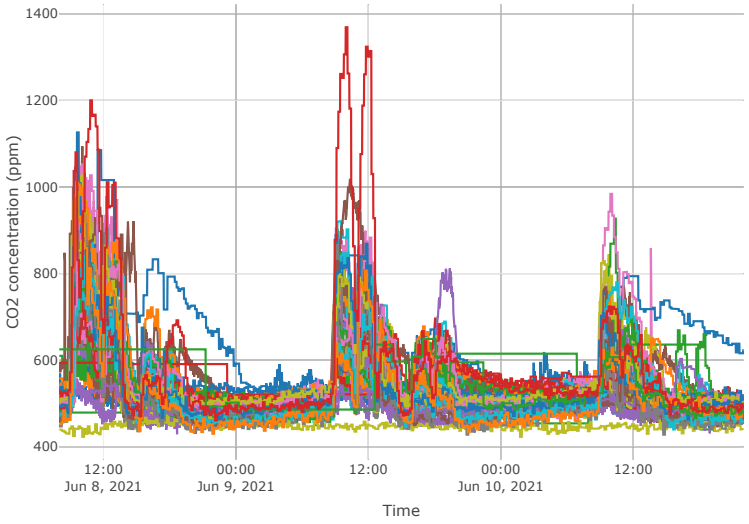



Figure 2: $CO_2$ levels for all days of EBAU

We can see that except for a few outliers the $CO_2$ levels tend to have a similar behaviour for all sensors. We can also see that the $CO_2$ levels tend to exhibit a periodic behaviour across the three days, with a very slight decreasing tendency, which is normal, as the exams with more enrolled students take place before those with less.

Now we will continue visualizing the functional data for the functions of an hour.

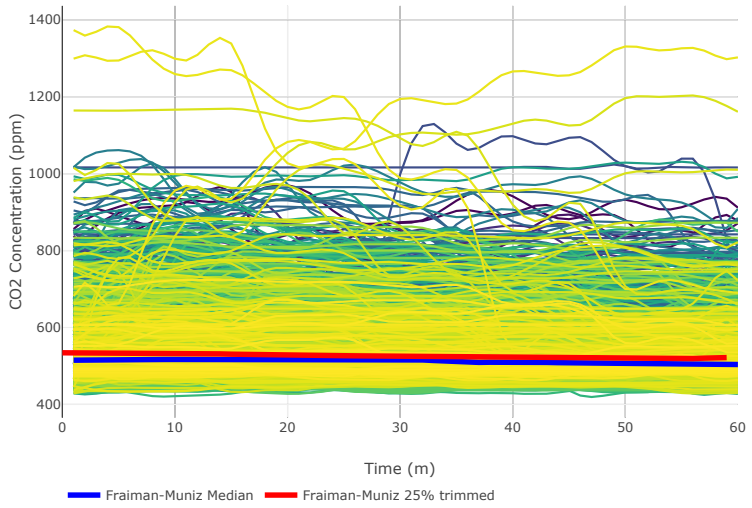



Figure 3: $CO_2$ Fraiman-Muniz depth for hourly $CO_2$ functions

In Figure 3 we see the Fraiman-Muniz depth for all the hourly $CO_2$ functions, which acts as a measure of central distribution. For this figure we utilised all the available curves, using functions for the different classrooms, days and hours. The functions with higher depth are coloured in yellow whereas those with lowest depth are coloured in dark blue. The deepest function, which can be interpreted as a median function is colored in blue and in red (calculated trimming 25% of the data). Observing this figure we see that altough there are some curves taking high levels of $CO_2$ and with high variations, the median takes safe levels of $CO_2$ and tends to be stable over the hour.

A functional ANOVA, as proposed in Cuevas et al., 2004, shows with a $p-$value of 0.46 that there is no statistical evidence to reject the null hypothesis of equality of the different faculties functional means.

## 4. Prediction of the mean $CO_2$ concentration level

In this section, we will focus on predicting the mean level of $CO_2$ reached in an hour given the information available before that hour. That is, given a filtration $\mathcal{F}_h$ that represents the information known up to the hour $h$, we will try to model the mean of the stochastic process in the instant $h+1$ adapted to that filtration: $\mathbb{E}[CO_{2h+1}|\mathcal{F}_h]$.

### 4.1. Simple model

We will begin with the simplest functional model we can build:

$$\mathbb{E}[CO_{2h+1}|\mathcal{F}_h] = \langle \beta_1(t), CO_{2h}(t) \rangle$$

That is, we assume that given all the information available to us up to hour $h$, the mean level of $CO_2$ reached in the subsequent hour is the inner product of some function of time $\beta_1(t)$ and the function of $CO_2$ at hour $h$.

Note here that we are making a prediction of the mean function of a real-valued process using a function-valued process. We predict the mean level of $CO_2$ in the next hour making use of the function of the current hour.

In order to fit this model, we estimate the sample mean of the scalar valued process at time $h$ as the mean of all the 60 observations we have for the different minutes at hour $h$. Then, we create sets of $B-spline$ basis to represent $CO_{2h}$ and $\beta_1(t)$ and we fit the model as in Ramsay and Silverman, 2005.

### 4.2. Variable selection model

We will now build another model with the variable selection algorithm proposed in Febrero-Bande et al., 2019 using the aforementioned variables together with these:

- $CO_2$ value taken in the $59^{th}$ minute of the current hour, $CO_{2h}(59)$.
- Mean and standard deviation of current hour $CO_2$ real-valued process.
- Fourier transforms and first derivatives of all our functional variables.

After applying the algorithm, we get the following Functional Spectral Additive Linear Model (GAM):

$$\mathbb{E}[CO_{2h+1}|\mathcal{F}_h] = s_1(CO_{2h}(59)) + s_2(h+1) + s_3(\sigma(CO_{2h}))+ \\ s_4(\mathbb{E}[CO_{2h}]) + s_5(CO_{2h}(t)) + s_6(\mathrm{Re}\{\mathrm{FT}(CO_{2h})\}) + s_7(\mathrm{Im}\{\mathrm{FT}(CO_{2h})\})$$

Where:

- $s_1, \ldots, s_7$ are smooth functions.
- $\mathrm{FT}(\cdot)$ denotes the Fourier transform of a function.
- $\mathrm{Re}\{\cdot\}, \mathrm{Im}\{\cdot\}$ denote the real and imaginary part of a complex number respectively.

### 4.3. Variable selection model with changes

To make another model, we tried to change some of the variables of the model built with a variable selection algorithm in order to obtain better prediction results. In this pursuit the following model was obtained:

$$\mathbb{E}[CO_{2h+1}|\mathcal{F}_h] = s_1(CO_{2h}(59)) + s_2(h+1) + s_3(\mathrm{WindSpeed}_h) + s_4(\mathrm{Re}\{\mathrm{FT}(CO_{2h})\})$$

where $s_1, \ldots, s_4$ are smooth functions.

We then obtain a model that predicts the mean level of $CO_2$ as a combination of smooth functions of the value taken in the last minute of the current hour, the next hour, the current wind speed and the real part of the Fourier transform of current hour $CO_2$ function.

### 4.4. Model Comparison

Table 2 collects the results of these fitted models. Recall that RMSE and MAE were calculated using 10-fold cross validation. The other metrics were computed using a model fitted with all available data.

We can see that although the second model improves the first one in terms of lower $AIC$ and higher $R^2_{adj}$, it performs worse in cross validation, resulting in higher $RMSE$ and $MAE$ and much higher standard deviation in these statistics. This could be caused by a tendency of the GAM model to overfit, so we tried to find another GAM model which corrects this effect.

We then obtain a model which has a better $AIC$ than the previous one and better $RMSE$, $MAE$ and standard deviations, giving us better results to justify its use instead of the simplest model.

| Model | RMSE | $\sigma_{RMSE}$ | MAE | $\sigma_{MAE}$ | AIC | $R^2_{adj}$ |
|---|---|---|---|---|---|---|
| 1 | 40.83 | 7.28 | 24.81 | 4.01 | 20254 | 79.38% |
| 2 | 57.87 | 35.23 | 25.28 | 4.99 | 19342 | 87.90% |
| 3 | 34.95 | 6.36 | 22.34 | 3.03 | 18626 | 86.70% |

Table 2: Results for mean prediction models

## 5 Conclusions

We have observed that models improving the simple model (which uses the current hour $CO_2$ curve as its only predictor) can be built. It is also to mention that, except for the wind speed, none of the ambient conditions variables (temperature, humidity and pressure) were selected by the variable selection algorithm and that their inclusion in the model doesn't seem to improve the model's results. Having obtaining as good results as we had, we conclude that this models can be utilized to make real time predictions with good accuracy.

**ACKNOWLEDGEMENTS**

This work, which is part of the CITIC's CEDCOVID project, has been supported by GAIN (Galician Innovation Agency) and the Regional Ministry of Economy, Employment and Industry, Xunta de Galicia grant COV20/00604 through the ERDF.

## REFERENCES

Allen, J. G., MacNaughton, P., Satish, U., Santanam, S., Vallarino, J., & Spengler, J. D. (2016). Associations of cognitive function scores with carbon dioxide, ventilation, and volatile organic compound exposures in office workers: A controlled exposure study of green and conventional office environments. *Environmental health perspectives*, *124*(6), 805–812.

Batterman, S. (2017). Review and extension of co2-based methods to determine ventilation rates with application to school classrooms. *International journal of environmental research and public health*, *14*(2), 145.

Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, *47*(1), 111–122.

Febrero-Bande, M., González-Manteiga, W., & de la Fuente, M. O. (2019). Variable selection in functional additive regression models. *Computational Statistics*, *34*(2), 469–487.

Peng, Z., & Jimenez, J. L. (2021). Exhaled co2 as a covid-19 infection risk proxy for different indoor environments and activities. *Environmental Science & Technology Letters*, *8*(5), 392–397.

Ramsay, J., & Silverman, B. (2005). *Functional data analysis. 2nd edition*. Springer-Verlag, New York.

Rudnick, S., & Milton, D. (2003). Risk of indoor airborne infection transmission estimated from carbon dioxide concentration. *Indoor air*, *13*(3), 237–245.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Relación entre las características sociodemográficas de las regiones europeas y la incidencia de COVID-19

Mª Isolina Santiago Pérez [1], Mª Esther López Vizcaíno [2], Cristina Candal-Pedreira[3], Mónica Pérez-Ríos[3] Alberto Ruano-Raviña[3]

[1] Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia

[2] Instituto Galego de Estatística

[3] Departamento de Medicina Preventiva y Salud Pública. Universidade de Santiago de Compostela

## RESUMEN

Europa ha sido uno de los continentes más afectados por la pandemia provocada por el COVID-19. El objetivo de este estudio es comparar la incidencia de COVID-19 entre los países y regiones de la Unión Europea (UE) que reportan datos al ECDC, a lo largo del período comprendido entre la semana epidemiológica 27 del año 2020 y la semana 53 del año 2020. También se evaluará si existe relación entre las características sociodemográficas de las regiones del territorio europeo y la incidencia de COVID-19.

**Palabras y frases clave:** COVID-19, incidencia, variabilidad, análisis funcional.

## 1. INTRODUCCIÓN

Europa ha sido uno de los continentes más afectados por la pandemia provocada por el COVID-19. A principios de marzo de 2021, se habían registrado 21.765.152 de casos de la enfermedad y 531.896 muertes por COVID-19 en Europa desde el inicio de la pandemia (1). Durante el periodo denominado primera ola, la mayor parte de países se vieron sorprendidos por la falta de preparación ante una crisis sanitaria sin precedentes en tiempos de paz en la historia europea. Esta crisis se manifestó a través de múltiples aspectos, como la escasez de test diagnósticos, la ausencia de equipos de protección individual para el personal sanitario y sociosanitario, la saturación de hospitales debido al elevado número de casos (2) y la falta de conocimiento sobre cómo cortar las cadenas de transmisión. Todo esto conllevó a un confinamiento generalizado en la mayoría de los países europeos. Tras la aplicación de medidas estrictas de Salud Pública, se observó que, a finales de junio de 2020, la incidencia en la mayoría de los países europeos había descendido significativamente. Sin embargo, a partir de octubre de 2020 la incidencia a nivel europeo volvió a aumentar, dando lugar a los períodos conocidos como segunda y tercera ola (3). A partir de octubre, estos aumentos en la incidencia no han sido homogéneos entre países, ni en cuanto a los picos de incidencia alcanzados ni en cuanto al momento en el que se han observado dichos picos. Así, mientras Francia y España tuvieron picos de incidencia a finales de octubre y principios de noviembre, el pico francés fue el doble del español, pero sin embargo Francia no sufrió un pico posterior, al contrario que España. Estas diferencias apuntan a un efecto debido a la distinta gestión de la pandemia en cada país europeo.

También se ha observado una importante variabilidad en la incidencia entre las regiones que conforman los países europeos (4) (5–7). Algunos autores han tratado de explicar los motivos de esta variabilidad proponiendo distintas causas, que se podrían clasificar como demográficas o asociadas a medidas de Salud Pública. Dentro de los factores demográficos destacan la estructura convivencial (intergeneracional o no), la densidad de población, la densidad de ocupación en las viviendas, el estilo de vida o costumbres de cada país y la frecuencia de uso de lugares de encuentro como bares o restaurantes. Adicionalmente, la edad media de la población juega un papel, al igual que la movilidad poblacional. Incluso los aspectos climáticos que podrían implicar más vida en espacios interiores podrían encuadrarse dentro de la vertiente demográfica. Por otro lado, la implementación de medidas de Salud Pública, como el confinamiento, el cierre de actividad económica no esencial o de las escuelas, el uso obligatorio de mascarilla, la implantación de toques de queda o el cierre de establecimientos de hostelería, se cree que han reducido la incidencia de COVID-19 significativamente (4,8,9).

Existe poca evidencia sobre la variabilidad de la incidencia de COVID-19 en Europa y las razones que pueden explicar este fenómeno. La aproximación formal a su análisis es compleja. Algunos autores se han centrado en la variabilidad, en la mortalidad y en la tasa de letalidad (7,11,12). Otros estudios han analizado cómo se comporta la incidencia de COVID-19 a nivel nacional (6,13). Un estudio ha analizado

la variabilidad en la incidencia en el continente europeo, pero no ha tenido en cuenta la variabilidad entre las regiones de un mismo país (5). Existen sin embargo metodologías relativamente sencillas que permiten aplicar una misma métrica a la variabilidad entre países e intrapaís en función de sus regiones, como por ejemplo el rango intercuartílico de variación en la incidencia entre regiones. Esta variación, aplicada a diferentes momentos (posteriores a la primera ola), en los que se ha obtenido evidencia y capacidad material acerca de cómo gestionar la pandemia, permite valorar el distinto desempeño de los países europeos y compararlos entre sí. Esto facilita conocer si la mayor o menor variabilidad entre regiones en la incidencia de COVID19 ocurre de forma puntual o si, por el contrario, existe una variabilidad sostenida en el tiempo que puede apuntar a fallos en la gestión por parte de la autoridad sanitaria.

El objetivo de este estudio es comparar la incidencia de COVID-19 entre los países de la Unión Europea (UE) que reportan datos al European Centre for Disease Prevention and Control (ECDC), además de comparar la incidencia de la enfermedad entre las distintas regiones de los países europeos a lo largo del período comprendido entre la semana epidemiológica 27 del año 2020 y la semana 53 del año 2020. También, se evalúa si existe relación entre las características sociodemográficas de las regiones del territorio europeo y la heterogeneidad de la incidencia de COVID-19.

## 2. MÉTODOS

### Fuentes de información

La fuente de información utilizada para los casos de COVID-19 es el ECDC. El día 10-03-2021 se descargaron de su página web (https://www.ecdc.europa.eu/en/covid-19/data) las tasas semanales a 14 días de casos COVID-19 por 100.000 habitantes de los 27 países de la Unión Europea (UE), a nivel nacional y a nivel regional, desde la semana 27 hasta la semana 53 del 2020. Para los datos de población de las regiones a 1 de enero de 2020 se utilizó EUROSTAT (http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_pjangrp3).

Los datos de las características sociodemográficas utilizadas se obtuvieron a nivel de región de EUROSTAT. Son correspondientes a los años 2018 o 2019.

Como criterios de inclusión se consideraron aquellas regiones que tenían información para la mayor parte de las variables sociodemográficas consideradas. Se analizaron datos de 177 regiones que pertenecen a 22 países distintos. Se excluyeron los países de Malta, Chipre, Luxemburgo, Estonia y Francia. Francia se tuvo que excluir del análisis porque en el ECDC solo tiene información a nivel de región a partir de la semana 53 del año 2020. Las regiones consideradas coinciden, en la mayor parte de los casos, con las NUTS2, a excepción de Bélgica y Alemania que coinciden con las NUTS1.

### Análisis estadístico

Con el propósito de que los países estuvieran en situaciones comparables con respecto a la epidemia del COVID-19, se empleó la información de las tasas a 14 días a partir de la semana 27 del año 2020, una vez superada la primera ola (29 de junio de 2020) hasta la semana 53 del año 2020, antes de que empezaran los efectos de la vacunación.

La información proporcionada por el ECDC a nivel de regiones tenía, para algunas semanas y regiones, valores negativos en las tasas de incidencia a 14 días. Para resolver esta situación, se eliminaron los valores negativos, y se imputaron esas tasas a partir del resto de datos empleando interpolaciones mediante splines (13).

Para cada país y cada una de las semanas consideradas, se calculó la mediana, el primer y tercer cuartil y el recorrido intercuartílico de las tasas regionales de incidencia a 14 días y se representaron en gráficos de evolución por país. El objetivo es identificar los países con mayor variabilidad entre las tasas de sus regiones.

Para cuantificar la variabilidad global de cada país, a lo largo de las 27 semanas del período analizado, se calcularon las distancias de las curvas regionales de tasas de incidencia a 14 días a la correspondiente curva nacional, utilizando para ello la métrica $L^2$ para datos funcionales (14,15). Se utiliza la siguiente expresión:

$$D\left(C_i(t), C_j(t)\right) = \sqrt[2]{\int \left(C_i(t) - C_j(t)\right)^2}$$

donde $C_{ij}(t)$ es la curva de la región i en el país j y $C_j(t)$ la curva del país j (t indica la semana).

Posteriormente, con las distancias regionales de cada país, se calculó un promedio ponderado por la población de las regiones, que sería un equivalente a la desviación típica, según la siguiente expresión:

$$d_j = \frac{\sum_{i=1}^{n_j} p_{ij}\, D\left(C_{ij}(t), C_j(t)\right)}{\sum_{i=1}^{n_i} p_{ij}}$$

donde $n_j$ es el número de regiones del país j y $p_{ij}$ es la población de la región i en el país j. El resultado, $d_j$, es una medida de variabilidad funcional del país j.

Por último, para facilitar la interpretación de la distancia ponderada, se estandarizaron los valores de todos los países a una escala entre 0 y 100 mediante la siguiente transformación:

$I_j = \frac{d_j - \min(d_j)}{\max(j) - \min(d_j)} * 100$ donde $d_j$ es la distancia ponderada del país j.

De este modo, el país con menor distancia o variabilidad tiene valor $I_j$=0, y el país con mayor distancia tiene valor $I_j$=100. El resto de países tienen un valor entre 0 y 100 que indica su posición relativa, respecto a la variabilidad entre el país con menor variabilidad (país 0) y el país con mayor variabilidad (país 100).

Para valorar la relación entre las características sociodemográficas de las regiones y la incidencia del COVID-19 se ajustó un modelo de regresión lineal mixto con efectos aleatorios. En este modelo la variable respuesta es la incidencia acumulada de las regiones del período de estudio de 27 semanas, que se aproximó sumando las tasas semanales de incidencia a 14 días dividido entre dos. Como variables explicativas se utilizaron las siguientes variables sociodemográficas:

- Densidad de población (habitantes/Km$^2$)
- Renta disponible de hogares privados (€)
- Número medio de personas por hogar
- % de personas en riesgo de pobreza o exclusión social
- Esperanza de vida al nacer
- Producto interior bruto (PIB) por habitante
- Tasa de desempleo (%)
- Edad mediana de la población
- Tasa de dependencia (población de 65 y más años entre la población de 15 a 64 años)
- Proporción de población de 15 años o menos
- Proporción de población de 15 a 64 años
- Proporción de población de 65 años y más
- Proporción de población de 80 años y más
- Proporción de personas en áreas densamente pobladas
- Pernoctaciones en alojamientos turísticos por 100.000 habitantes
- Porcentaje de ocupados que trabajan en el sector primario
- Porcentaje de ocupados que trabajan en el sector secundario
- Región costera (0-sin costa, 1-con costa)
- Región insular (0-isla, 1-no isla)

Algunas de las variables explicativas tenían datos faltantes para algunas de las regiones. Cuando se producía esta situación se le imputó el dato del país. Es el caso, por ejemplo, de la variable tasa de riesgo de pobreza y exclusión social que no está disponible para las regiones de Alemania.

Para descartar aquellas variables que no influyen significativamente en la incidencia de COVID-19, en un primer momento se ajustó un modelo de regresión lineal por pasos y con las variables seleccionadas en el paso anterior, se ajustó el modelo de regresión lineal mixto con efectos aleatorios. Los coeficientes de las variables sociodemográficas son un factor fijo dentro del modelo y el país es un factor aleatorio. Una vez ajustado el modelo se validó mediante el estudio de los residuos.

## 3. RESULTADOS

Desde el principio de la pandemia hasta el 31 de diciembre de 2020 se diagnosticaron aproximadamente 27,2 millones de casos de COVID-19 en Europa, lo que supone una incidencia acumulada de casi 3.209 casos por 100.000 habitantes. Esta tasa varía entre 668 de Finlandia y los 10.700 de Andorra. En los 22 países de la UE incluidos en este análisis la incidencia acumulada fue de 3.489 casos

por 100.000 habitantes (13,14 millones de casos). Por países, la incidencia mínima se observó en Finlandia y la máxima en la República Checa, con 668 y 6.982 casos por 100.000 habitantes, respectivamente.

La Figura 1 (izquierda) presenta en un mapa la incidencia acumulada en las regiones del estudio. Las regiones con colores más oscuros se corresponden con zonas donde la incidencia es mayor y las zonas más claras reflejan incidencias menores. Se observa una mayor incidencia en regiones del centro de España y en un arco que va desde el norte de Italia a Polonia, pasando por Austria y República Checa, además de Croacia. Regiones de Bélgica, Holanda y Lituania también presentan altas tasas. Además, también se observa una cierta correlación espacial que fue corroborada con el test de Moran ($p<0.05$). En la Figura 1 (derecha) se presenta la variabilidad de la incidencia acumulada en las regiones de los 22 países. Se observan las mayores tasas en República Checa y Eslovenia y las menores en Finlandia. Se muestran, además, diferentes variabilidades entre países, los de mayor variabilidad son España, Italia, Bélgica y Portugal.

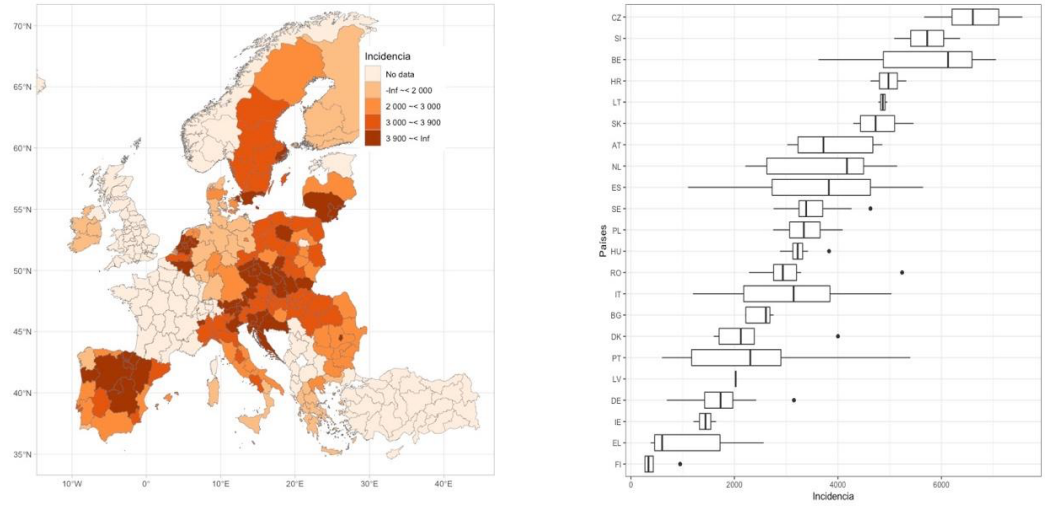



Figura 1. Incidencia acumulada en las regiones de los 22 países europeos incluidos, desde la semana 27 hasta la 53 de 2020

En la figura 2 se representa la evolución semanal de la tasa mediana de incidencia a 14 días por país, y su recorrido intercuartílico desde la semana 27 de 2020. Se observa un comportamiento diferente de la pandemia entre países, tanto en el nivel de incidencia como en la variabilidad entre regiones. La mayor variabilidad parece observarse en Bélgica, Austria, España y Portugal. Puede observarse en la figura que la incidencia hasta el 7 de septiembre es muy baja (ver línea vertical) en todos los países analizados excepto España, donde la incidencia comienza a subir en el mes de julio.

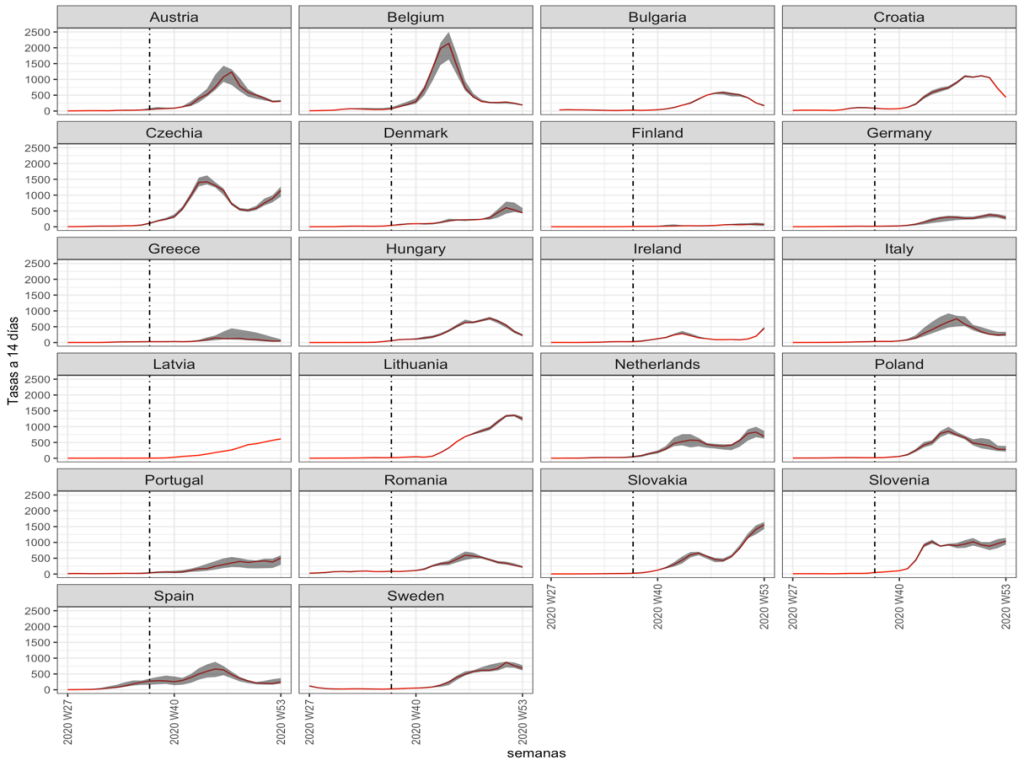



Figura 2. - Tasas de incidencia a 14 días de COVID-19 en los 22 países de la UE incluidos, desde la semana 27 de hasta la semana 53 de 2020. Mediana de las tasas regionales de cada país (línea roja) y recorrido intercuartílico (banda gris).

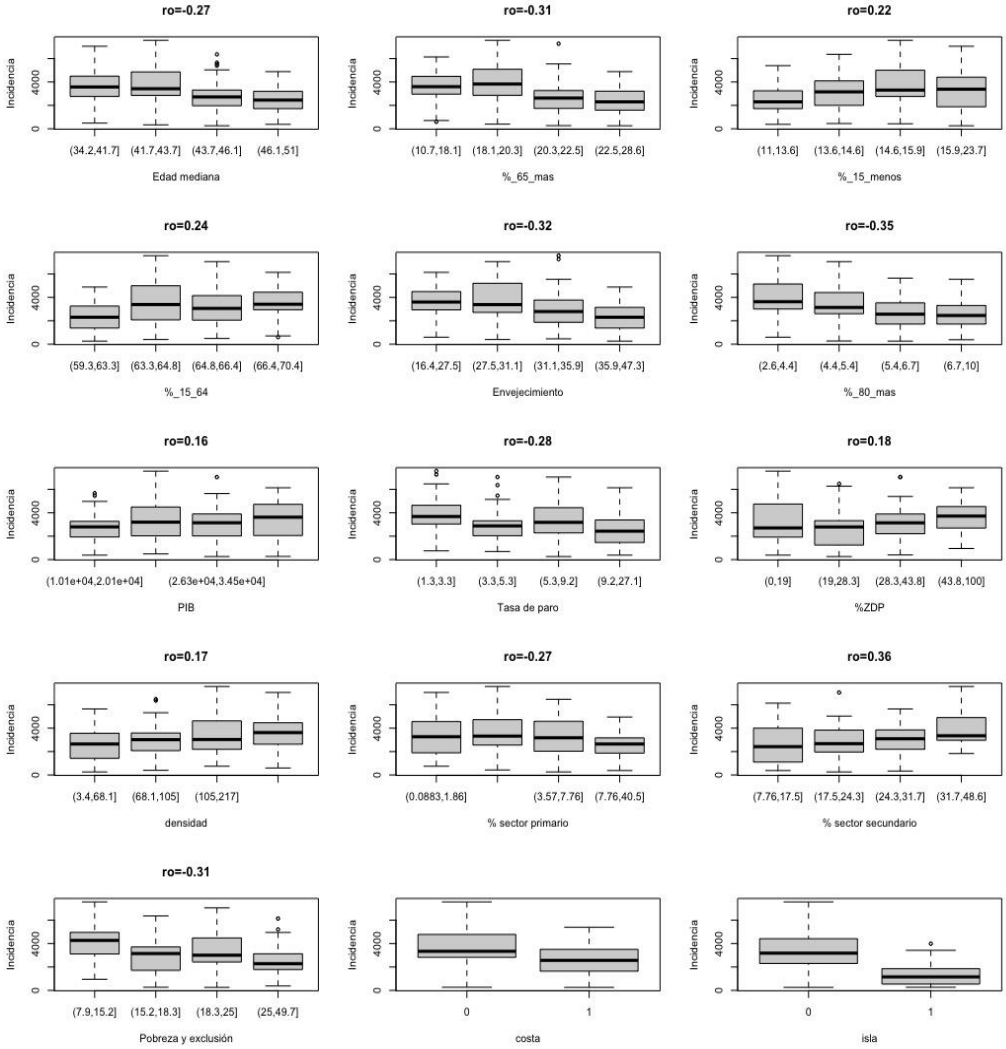
Si se analiza la variabilidad entre las curvas de incidencia durante el período de 27 semanas (Tabla 1), Bélgica, Portugal y España son los tres países que ocupan las primeras posiciones en cuanto a la variabilidad funcional.

Tabla 1. Variabilidad funcional de cada país, ponderada por la población de las regiones, empleando la métrica $L^2$. Países ordenados de mayor a menor variabilidad.

| País | Variabilidad funcional *(d)* | *I (escalada de 0 a 100)* |
|---|---|---|
| Belgium | 1304,9 | 100 |
| Portugal | 865,1 | 66 |
| Spain | 703,2 | 53 |
| Slovenia | 703,2 | 53 |
| Italy | 668,6 | 50 |
| Czechia | 614,0 | 46 |
| Austria | 600,3 | 45 |
| Denmark | 571,4 | 43 |
| Slovakia | 522,3 | 39 |
| Netherlands | 488,3 | 36 |
| Greece | 484,4 | 36 |
| Sweden | 482,0 | 36 |
| Poland | 471,3 | 35 |
| Romania | 429,7 | 32 |
| Croatia | 322,5 | 23 |
| Bulgaria | 317,9 | 23 |
| Hungary | 264,4 | 19 |
| Germany | 252,1 | 18 |
| Lithuania | 198,4 | 14 |
| Finland | 146,0 | 10 |
| Ireland | 115,3 | 7 |
| Latvia | 23,3 | 0 |

En la Figura 3 se presenta la relación entre la incidencia acumulada y las variables explicativas categorizadas en cuartiles mediante boxplots. También se presenta el coeficiente de correlación por rangos de Spearman (ro). Se puede observar como, a medida que aumenta la mediana de edad en las regiones, el porcentaje de población de 65 y más años, la tasa de paro o la tasa de pobreza y exclusión social, también disminuye la incidencia acumulada. Por el contrario, si aumenta el PIB, el porcentaje de población en zonas densamente pobladas, el porcentaje de población de 15 y menos años, la densidad de población o el porcentaje de ocupados en el sector secundario, aumenta la incidencia acumulada. También se puede observar como el hecho de que una región tenga costa o que sea una isla, influye en la incidencia de COVID-19.

Figura 3.- Relación entre la incidencia acumulada en las regiones de los 22 países de la UE desde la semana 27 hasta la 53 de 2020 y las variables sociodemográficas




El ajuste del modelo de regresión lineal aporta un $R^2$ de 0,80, siendo significativas las variables: % población de 65 y más años, proporción de personas en áreas densamente pobladas, % de personas en riesgo de pobreza o exclusión social, porcentaje de ocupados que trabajan en el sector secundario, indicador de costa e indicador de región insular. La Tabla 2 muestra los resultados del ajuste del modelo lineal mixto con efecto aleatorio de país, los coeficientes de regresión de las variables asociadas con el porcentaje de área bajo la curva junto con el error estándar y el p-valor. Como se muestra, un mayor porcentaje de población en zonas densamente pobladas y un mayor porcentaje de ocupados en el sector secundario (zonas más industrializadas) está asociada con regiones de más incidencia de COVID-19, mientras que la mayor proporción de población de 65 y más años y la mayor tasa de riesgo de pobreza y exclusión social está asociada con regiones con menor incidencia de COVID-19. Además, las regiones con costa y las islas también están asociadas con menores incidencias.

| | Coeficientes | Error estándar | t-valor | p-valor |
|---|---|---|---|---|
| % población de 65 y más años | -84,04 | 27,48 | -3,06 | 0,00 |
| Porcentaje de población en zonas densamente pobladas | 12,59 | 3,31 | 3,80 | 0,00 |
| % de personas en riesgo de pobreza o exclusión social | -31,72 | 8,75 | -3,62 | 0,00 |
| % de ocupados que trabajan en el sector secundario | 42,54 | 11,14 | 3,82 | 0,00 |
| Región con costa | -499,75 | 149,77 | -3,34 | 0,00 |
| Región insular | -835,72 | 255,14 | -3,28 | 0,00 |

Tabla 2. Resultados del ajuste de modelo de regresión lineal mixto con efectos aleatorios

Se validó el modelo anterior empleando para ello el gráfico qq-plot (Figura 4 izquierda) y el gráfico de residuos estandarizados frente a los valores predichos (Figura 4 derecha). En el gráfico qq-plot se observa que los puntos están próximos a la diagonal, excepto los tres últimos, que se corresponden con la Región Norte en Portugal, Hovedstaten en Alemania y Region Wallone en Bélgica. Todas ellas son regiones que alcanzan valores muy altos de incidencia si las ponemos en referencia al resto de las regiones de sus países. En el gráfico de los residuos estandarizados no se observa ningún patrón. También se evaluó la correlación espacial de los residuos utilizando el test de Moran y no se observó patrón espacial (p=0,30).

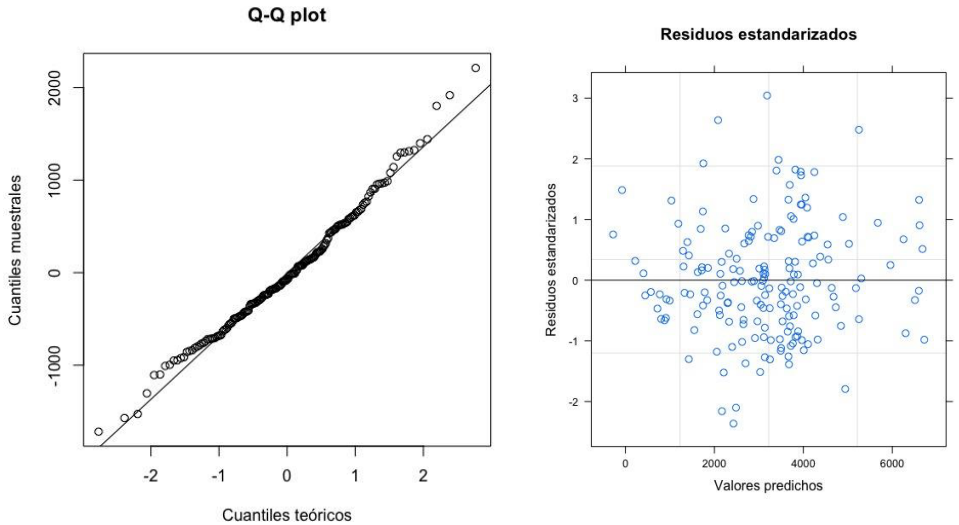



Figura 4.- Q-Q plot y gráfico de los residuos estandarizados frente a los valores predichos

## 3. CONCLUSIONES

Podemos concluir que la variabilidad de la incidencia por COVID-19 entre países europeos, y también dentro de un mismo país, ha sido muy importante tras la primera ola. Estas diferencias podrían explicarse por fenómenos de transmisión diferenciales entre países, por factores sociodemográficos, pero también

podrían deberse a cómo los Gobiernos han encarado la situación epidemiológica una vez que se llegó a un control epidemiológico relativo a comienzos del verano de 2020. Ha habido grandes diferencias en la incidencia admisible por cada Gobierno a nivel de los distintos países europeos, lo que limita en gran medida la variabilidad en la incidencia al establecerse un límite superior bajo.

En lo referente a los factores sociodemográficos, nuestros resultados parecen demostrar una relación entre la incidencia del COVID-19 y las zonas densamente pobladas, lo que reforzaría la importancia del aislamiento social en el control de la pandemia. Respecto a la menor incidencia observada en las regiones con mayores porcentajes de población mayor de 65 años podría deberse a que este grupo poblacional ha adoptado mayores precauciones para evitar el contagio y ha limitado su interacción social, de partida ya más limitada que en grupos de edad inferiores. Además, la identificación de características sociodemográficas regionales asociadas con una incidencia más alta de COVID-19 podría ayudar a las autoridades de salud a realizar una asignación más racional de los recursos disponibles y a combatir de manera más efectiva la pandemia.

## REFERENCIAS

1. Weekly Reports. ECDC. Week 07, 2021 [Internet]. [citado 2021 Mar 3]. Disponible en: https://covid19-surveillance-report.ecdc.europa.eu/
2. Han E, Tan MMJ, Turk E, Sridhar D, Leung GM, Shibuya K, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. The Lancet. 2020;396(10261):1525–34.
3. WHO Coronavirus (COVID-19) Dashboard [Internet]. [citado 2021 Mar 5]. Disponible en: https://covid19.who.int
4. Alfano V, Ercolano S. The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel Analysis. Appl Health Econ Health Policy. 2020;18(4):509–17.
5. Srivastava A, Chowell G. Understanding Spatial Heterogeneity of COVID-19 Pandemic Using Shape Analysis of Growth Rate Curves. medRxiv [Internet]. 2020 [citado 2021 Mar 3]; Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273268/
6. Santiago Pérez MI, López-Vizcaíno E, Ruano-Ravina A, Pérez-Ríos M. Sistema de ayuda a la toma de decisiones sanitarias. Propuesta de umbrales de riesgo epidemiológico ante SARS-CoV-2. Arch Bronconeumol [Internet]. 2021[citado 2021 Mar 3]; Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7826127/
7. Sorci G, Faivre B, Morand S. Explaining among-country variation in COVID-19 case fatality rate. Sci Rep [Internet]. 2020[citado 2021 Mar 3];10. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7609641/
8. Wong CKH, Wong JYH, Tang EHM, Au CH, Lau KTK, Wai AKC. Impact of National Containment Measures on Decelerating the Increase in Daily New Cases of COVID-19 in 54 Countries and 4 Epicenters of the Pandemic: Comparative Observational Study. J Med Internet Res. 2020;22(7):e19904.
9. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong Social Distancing Measures In The United States Reduced The COVID-19 Growth Rate. Health Aff Proj Hope. 2020;39(7):1237–46.
10. Zhang X, Warner ME. COVID-19 Policy Differences across US States: Shutdowns, Reopening, and Mask Mandates. Int J Environ Res Public Health. 2020;17(24).
11. Hradsky O, Komarek A. Demographic and public health characteristics explain large part of variability in COVID-19 mortality across countries. Eur J Public Health. 2021;31(1):12–6.
12. Miller LE, Bhattacharyya R, Miller AL. Data regarding country-specific variability in Covid-19 prevalence, incidence, and case fatality rate. Data Brief. 2020;32: 106276.
13. Forsythe, G. E., Malcolm, M. A. and Moler, C. B. (1977). Computer Methods for Mathematical Computations. Wiley.
14. Silverman, B.W. and Ramsay, J.O (2005). Functional Data Analysis. Springer, Second Edition.

15. Febrero-Bande, M., Oviedo de la Fuente, M (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51(4), 1-28. Disponible en http://www.jstatsoft.org/v51/i04/

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# UNHA NOVA HEURÍSTICA EN DÚAS FASES PARA UN PROBLEMA DE REPARTO DE PENSOS CON CAMIÓNS E REMOLQUES DIVIDIDOS EN COMPARTIMENTOS

Laura Davila-Pena[1], David R. Penas[1] e Balbina Casas-Méndez[1]

[1]Grupo de Modelos de Optimización, Decisión, Estatística e Aplicacións (MODESTYA), Departamento de Estatística, Análise Matemática e Optimización, Instituto de Matemáticas (IMAT), Facultade de Matemáticas, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, España

## RESUMO

Neste traballo presentamos un novo modelo para o problema de rutas no que a frota dispoñible está formada por camións e remolques divididos en compartimentos, ao que chamamos *multi-compartment truck and trailer routing problem* (MC-TTRP). Este problema está motivado polas necesidades dunha cooperativa agrícola galega que distribúe penso para o gando (Guitián de Frutos e Casas-Méndez, 2019). Unha formulación tentativa do modelo xa foi introducida nun traballo preliminar de Davila-Pena (2019). Dado que o MC-TTRP é un problema NP-duro, resolver o modelo mediante métodos exactos para problemas grandes é moi custoso computacionalmente. Deste xeito, introducimos e implementamos un algoritmo heurístico en dúas fases. Na primera fase, obtense unha solución inicial a partir dunha heurística construtiva baseada no algoritmo dos aforros de Clarke e Wright (1964). Na segunda fase, mellórase esta solución inicial mediante unha búsqueda tabú iterativa. Esta última metaheurística está inspirada en varios traballos relacionados, tales como o de Cordeau e Maischberger (2012) para algunhas variantes do *vehicle routing problem* (VRP) ou o método implementado por Silvestrin e Ritt (2017) para o MC-VRP.
Este algoritmo foi probado en 21 instancias clásicas do TTRP (Chao, 2002), así como en 21 problemas xerados para o MC-TTRP a partir destas últimas. Os resultados do noso estudo computacional proban a efectividade da nosa proposta, acadando sempre solucións factibles de boa calidade. Preséntase tamén unha aplicación do modelo e a heurística propostos ao caso da cooperativa agrícola galega mencionada anteriormente, comparando as solucións obtidas mediante as dúas metodoloxías.
En Davila-Pena et al. (2021) móstranse estes resultados de forma detallada. Ademais, o código de dita heurística así coma os datos empregados para validala están presentes no repositorio de GitHub creado a tal efecto:

https://github.com/LauraDavilaPena/ITS_MC-TTRP.

**Palabras e frases chave**: problemas de rutas con camións e remolques; vehículos compartimentados; algoritmo heurístico construtivo; búsqueda tabú; loxística.

## REFERENCIAS

Chao, I. M. (2002) A tabu search method for the truck and trailer routing problem. Computers & Operations Research, 29, 33–51.

Cordeau, J. F. and Maischberger, M. (2012) A parallel iterated tabu search heuristic for vehicle routing problems. Computers & Operations Research, 39, 2033–2050.

Davila-Pena, L. (2019) Modelos y algoritmos en una clase de problemas de rutas de vehículos. Master's thesis, Universidade de Santiago de Compostela. Dispoñible en:
http://eamo.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1685.pdf.

Davila-Pena, L., R. Penas, D. and Casas-Méndez, B. (2021) A new two-phase heuristic for the routing problem of food distribution with compartmentalized trucks and trailers. Under revision.

Guitián de Frutos, R. M. and Casas-Méndez, B. (2019) Routing problems in agricultural cooperatives: a model for optimization of transport vehicle logistics. IMA Journal of Management Mathematics, 30, 387–412.

Silvestrin, P. V. and Ritt, M. (2017) An iterated tabu search for the multi-compartment vehicle routing problem. Computers & Operations Research, 81, 192–202.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Planificación de tarefas nun servizo hospitalario de quimioterapia apoiada nun modelo de programación estocástica

Adrián González Maestro[1], Elena Brozos Vázquez[2], Balbina Casas Méndez[3], Rafael López López[2], Rosa López Rodríguez[2] e Francisco Reyes Santias[4]

[1] Investigador. Fundación IDIS, Santiago de Compostela.
[2] Servicio de Oncoloxía. Hospital Clínico Universitario de Santiago de Compostela.
[3] Dpto. Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela.
[4] Economista. Fundación IDIS, Santiago de Compostela.

## RESUMO

Neste traballo proponse un modelo lineal de programación estocástica con variables enteiras, encamiñado á planificación do horario dos tratamentos dos pacientes nun servizo hospitalario de Oncoloxía Médica. O obxectivo do modelo é minimizar a duración das estancias na sala de espera do conxunto total dos pacientes. O modelo permite ademáis reorganizar os horarios de citas médicas cos oncólogos e compleméntase cunha ferramenta que resolve o problema de asignación de enfermeiras a pacientes. O traballo vén motivado polas características particulares dun caso real e o modelo utilízase e compárase con datos do devandito caso real.

**Palabras e frases chave:** asignación de enfermeiras, citas médicas, estudo dun caso, horarios de tratamentos, programación estocástica, programación lineal enteira.

## 1. INTRODUCIÓN

Na actualidade, unha gran parte dos casos diagnosticados de cancro son tratados sen necesidade de realizar unha hospitalización do paciente, o cal en xeral resulta ser unha gran axuda que favorece o benestar do paciente e unha mellor calidade de vida. Normalmente, as terapias oncolóxicas lévanse a cabo en centros de día aos que o paciente acode para efectuar os distintos requirimentos do seu tratamento, regresando despois ao seu domicilio, onde ten lugar a recuperación da terapia.

Por outra banda, durante as últimas décadas, o número de casos de pacientes con cancro subiu significativamente (incremento que seguimos experimentando hoxe en día) debido principalmente ao aumento da expectativa de vida dos individuos. Son precisamente as persoas de máis avanzada idade as máis propensas a padecer estas enfermidades. Unha maior demanda implica unha maior dificultade á hora de manter a calidade dos servizos ofrecidos aos pacientes. Concretamente, referímonos á capacidade de manter estables os tempos de espera dos pacientes durante a súa estancia nos devanditos centros.

Cabe sinalar que este tipo de centros teñen uns protocolos de acción bastante complexos que poden ser propensos a xerar importantes tempos de espera para os pacientes. Hai que ter en conta que se trata de servizos de alta demanda e que se combinan cunha organización que en ocasións presenta marxes de mellora. Todo isto implica que a planificación dos centros clínicos destinados a tratar a pacientes con cancro sexa unha tarefa relevante, que preocupa aos diversos profesionais implicados en hospitais e clínicas tanto da nosa contorna máis próxima, de Galicia, como a nivel nacional e internacional. É de destacar que os tratamentos existentes contra esta enfermidade son bastante caros, e por tanto, fronte á obvia (pero en ocasións inviable) solución de investir máis diñeiro nos devanditos centros, xorde a cuestión de se é posible unha mellora da organización dos distintos procedementos que levan a cabo nestas unidades, sen aumentar necesariamente os recursos dos que se dispón.

Todo o proceso deseñado para a atención e tratamento dos pacientes oncolóxicos, conformado por diversas etapas a cubrir polo paciente o día que acode ao hospital, involucra a colaboración de moitos profesionais do equipo, especializados en tarefas moi diferentes (enfermeiras, médicos, farmacéuticos, etc.), e o emprego de certos recursos limitados, tanto materiais (salas de citas, cadeiras especiais para tratamentos, etc.) como humanos. Por tanto, a existencia dun protocolo de actuación eficiente entre todos estes axentes resulta ser un ingrediente capital á hora de minimizar o tempo de espera dos pacientes, a creación das diferentes citas e a planificación de persoal de enfermería e outros profesionais implicados.

## 2. O PROBLEMA DO SERVIZO HOSPITALARIO DE ONCOLOXÍA

A continuación, profundizamos na descrición do problema tratado nesta comunicación e explícase a metodoloxía abordada. Ademáis, preséntanse algúns traballos recentes da literatura relacionada e dáse conta da proposta de solución adoptada.

Cabe mencionar que todo o traballo ten a súa orixe nas prácticas realizadas polo primeiro autor deste traballo no hospital universitario de Santiago de Compostela, para a realización do seu traballo fin de máster ao amparo do convenio existente entre o máster interuniversitario en técnicas estatísticas, das tres universidades galegas, e o SERGAS, o Servizo Galego de Saúde.

### 2.1. DESCRICIÓN DO PROBLEMA

No Hospital de Día de Oncoloxía (HDO) de Santiago de Compostela aténdese cada xornada a pacientes de cancro ambulatoriamente proporcionando dous tipos de servizos médicos, sendo estos as consultas cos facultativos e os tratamentos de quimioterapia. Ditas actuacións están fortemente relacionados, sendo a consulta, isto é, o acto médico, un antecedente necesario para poder levar a cabo os tratamentos máis adiante, aínda que durante a mesma xornada.

A forma de proceder actual do hospital consiste en darlle aos pacientes unha estimación do horario das súas citas, deixando o comezo do tratamento como un horario indeterminado que se lle dará a coñecer ao paciente só nos momentos previos a que éste tome lugar. Isto xera a problemática de que o paciente debe permanecer na sala de espera do hospital entre ambos servizos médicos, o que en moitas ocasións provoca, como mínimo, incerteza, e, ás veces, malestar aos pacientes.

### 2.2. METODOLOXÍA ADOPTADA

Plantexámonos usar a optimización matemática para obter un algoritmo que, proporcionando uns horarios para os pacientes completamente definidos, (é dicir, que incluan unha estimación tanto da revisión oncolóxica como do tratamento de quimioterapia), minimice ademais os tempos de agarda entre ambos servizos médicos.

Con tal fin, organizáronse entrevistas con todos os profesionais implicados no circuito que experimentan os pacientes do HDO e consultouse a bibliografía existente acerca do tópico da optimización matemática aplicada á quimioterapia ambulatoria.

O noso obxectivo era ter unha comprensión global de como adoitan ser atacados matemáticamente este tipo de problemas e, a partir dese coñecemento, poder ofrecer unha solución que sexa válida para as circunstancias concretas do HDO.

Durante a semana do 11 de xaneiro de 2021 recollimos datos no HDO de Santiago de Compostela acerca da duración dos distintos tempos relativos ás consultas cos oncólogos e os tratamentos de quimioterapia.

Concretamente, tomáronse os tempos das diferenzas horarias entre as horas teóricas e reais dos comezos das revisións oncolóxicas, os tempos das duracións de ditas revisións, os tempos das diferenzas horarias entre as finalizacións das revisións e os momentos nos que as substancias están listas para ser administradas, e finalmente os tempos das duracións dos tratamentos de quimioterapia.

A partir destes datos podemos facernos unha idea de onde se atopan as etapas críticas do proceso e de cómo afrontar o posible modelado do problema.

### 2.3. LITERATURA RELACIONADA

Existen diversos traballos abordando as labores de optimizar a programación de citas de centros de día de oncoloxía. En Turkcan et al. (2012) preséntase a situación dun centro clínico ao cal os pacientes acuden para recibir sesións de quimioterapia cunha certa periodicidade, formando ciclos regulares de sesións de tratamento. Perséguense dous obxectivos. Por un lado, reducir o máximo posible os retrasos nos ciclos dos pacientes, dado que as dilatacións execesivas dos lapsos de tempo entre sesións de quimioterapia diminúen en gran medida a súa efectividade. Por outro lado, reducir os custos do hospital asociados ás horas de traballo que son levadas a cabo.

En Liang et al. (2015) abórdase o problema de minimizar os tempos de espera dos pacientes durante un día de traballo nun centro onde ofrécense citas con oncólogos e tratamentos de quimioterapia nun contexto no que xa é sabido cantos pacientes hai que atender e de que tipo son. Inicialmente, plantéxase un problema de optimización bi-obxectivo, procurando acadar un traballo balanceado ao longo das xornadas laborais tanto no referente ao uso das cadeiras de quimioterapia como no referente ás consultas cos oncólogos.

En Heshmat e Eltawil (2021), no contexto dun centro ao cal acuden os pacientes para recibir quimioterapia, abórdase o problema de minimizar os retrasos nos tratamentos dos pacientes e o tempo de traballo total do centro.

O modelado do artigo de Hesaraki et al. (2019) pode aplicarse en centros de día oncolóxicos cunhas necesidades bastante concretas. En dito documento estúdase unha forma de proceder cando, para un día de traballo dado, os pacientes do centro teñen asignadas citas co seu oncólogo e tratamentos de quimioterapia. Ademáis, pártese da hipótese de que as citas cos oncólogos de cada un dos pacientes están programadas de antemán, e o problema enfrentado consiste en determinar os horarios das sesións de tratamento de quimioterapia de cada paciente. Así pois, este é un traballo próximo á problemática que nos ocupa, aínda que as restricións e o propio obxectivo non son idénticos aos do HDO. Con todo, é destacable que o traballo presenta a gran vantaxe de que o modelo proposto pode resolverse de maneira exacta de forma moi rápida, para a planificación dun día de traballo cun número de pacientes ao redor de 100.

### 2.4. PROPOSTA DE SOLUCIÓN

Cabe mencionar que nos traballos citados, como na maioría dos atopados na literatura, faise uso de modelos de programación matemática de tipo determinístico. Con todo, a incerteza existente nos tempos das distintas etapas involucradas no proceso fixeron aconsellable a proposta dun modelo de tipo estocástico que represente de maneira máis adecuada a realidade presente, aínda que entraña unha maior complexidade no modelado e implica uns tempos máis longos para a obtención de solucións exactas, ou mesmo unicamente aproximadas cando se contemplan amplos horizontes de tempo (semanais ou mesmo mensuais). Os detalles do modelado proposto e outra información relevante pódese consultar en González Maestro (2021).

## 3. CONCLUSIÓNS

Nesta comunicación, preséntase un modelo de programación estocástica en dúas etapas (Birge e Louveaux, 2011). Asumimos a incerteza respecto do tempo de duración das distintas etapas do circuíto dos pacientes e consideramos varios posibles escenarios para a fluidez do conxunto de pacientes a través das distintas etapas do proceso. Para o deseño destes escenarios fixéronse uso de técnicas de clasificación.

Expomos así un modelo que ofrece varios posibles horarios para a quimioterapia de cada paciente. A función dos profesionais do hospital de día oncolóxico é decidir, segundo como vaia transcorrendo a xornada laboral, qué estimación comunicarlle a cada paciente para o comezo do tratamento. Ademais, neste modelo introducimos tamén a posibilidade de poder replanificar as revisións oncolóxicas para facer que a minimización das esperas dos pacientes entre compromisos médicos sexa mais efectiva. Por suposto, dita planificación de horarios das revisións respecta unhas restriccións mínimas que aseguran que os horarios son asumibles polos oncólogos encargados da primeira etapa do proceso.

A maiores de conseguir uns horarios definidos para os pacientes respecto aos dous compromisos médicos que teñen, este modelo proporciona unha mellora estimada dos tempos de espera do 17%, medida obtida a partir dos datos dispoñibles. O modelo estocástico, compleméntase cunha ferramenta que resolve o problema de asignación de enfermeiras a pacientes, que fai uso, como parámetros, dos resultados do modelo estocástico.

Como é propio desta metodoloxía, os modelos propician unha análise de post-optimalidade que permitiría ver o efecto de modificacións nos parámetros do modelo. Notemos que facemos uso de parámetros deterministas (fixos) como número de enfermeiras, farmacéuticos, médicos, ou tamén salas de consulta, ou cadeiras de quimioterapia. Tamén contamos con parámetros estocásticos (variables): as distribucións dos tempos das diversas etapas do circuíto.

No tocante aos aspectos computacionais do traballo, debemos mencionar que todos os modelos estudados foron resoltos coa linguaxe de programacion matemática AMPL (Fourer et al., 2003) e facendo uso do solucionador Gurobi (https://www.gurobi.com/). Os tempos de computación, considerando un horizonte de tempo correspondente a unha xornada de traballo, nunca superaban os 30 segundos de duración (aspecto importante para aplicar o modelo no ámbito sanitario).

## AGRADECEMENTOS

Moitas grazas a Beatriz, Nieves e todos os profesionais sanitarios do CHUS que coa súa colaboración fixeron que fose posible levar a cabo este traballo.

## REFERENCIAS

Birge, J. R. and Louveaux, F. (2011). Introduction to Stochastic Programming. Second Edition. Springer Series in Operations Research and Financial Engineering.

Fourer, R., Gay, D. M., and Kernighan, B. W. (2003). AMPL. A Modeling Language for Mathematical Programming. Second Edition. Duxbury Thomson.

González Maestro, A. (2021). Optimización dos circuitos de pacientes do Hospital de Día de Oncoloxía. Traballo Fin de Máster. Máster en Técnicas Estatísticas. Universidade de Santiago de Compostela. Consulta online (18/07/21): http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1909.pdf

Hesaraki, A. F., Dellaert, N. P., and de Kok, T. (2019). Generating outpatient chemotherapy appointment templates with balanced flowtime and makespan. European Journal of Operational Research, 275 (1), 304-318.

Heshmat, M. and Eltawil, A. (2021). Solving operational problems in outpatient chemotherapy clinics using mathematical programming and simulation. Annals of Operations Research, 298, 289-306.

Liang, B., Turkcan, A., Ceyhan, M. E., and Stuart, K. (2015). Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. International Journal of Production Research, 53 (24), 7177-7190.

Turkcan, A., Zeng, B., and Lawley, M. (2012). Chemotherapy operations planning and scheduling. IIE Transactions on Healthcare Systems Engineering, 2 (1), 31-49.

.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

## Global optimization for bilevel portfolio design: Economic insights from the Dow Jones index

Julio González-Díaz [1,2,3], Brais González-Rodríguez [1,2], Marina Leal [4,5] Justo Puerto [5]

[1]Department of Statistics, Mathematical Analysis and Optimization and IMAT, University of Santiago de Compostela, Spain
[2]MODESTYA Research Group, Santiago de Compostela, Spain
[3]ITMATI (Technological Institute for Industrial Mathematics), Santiago de Compostela, Spain
[4]ALOP, Trier University, Germany
[5]IMUS, University of Seville, Spain

### RESUMEN

En este trabajo se presenta un problema de selección de carteras con costes de transacción y dos niveles de decisión. Se considera que hay un bróker que controla las comisiones que se cobran por los distintos activos, con el objetivo de maximizar su beneficio. Por otra parte, hay un inversor que elige su cartera tratando de minimizar el riesgo, garantizándose al mismo tiempo un rendimiento mínimo. Esto da lugar a una estructura jerárquica y se presentarán diferentes modelos de programación matemática para las diferentes situaciones, dependiendo de quién sea el primero en la jerarquía. Finalmente, se mostrarán los resultados obtenidos tras un análisis computacional con los datos del índice Dow Jones.

**Palabras y frases clave**: Portfolio design, Nonlinear programming, Bilevel optimization, Global optimization

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

**Detección automática de norias y de sus zonas de carga y descarga**

Manuel Antonio Novo Pérez[1], Marta Rodríguez Barreiro[2], Manuel Vaamonde Rivas[1] y María José Ginzo Villamayor[3]

[1]Universidade da Coruña, Facultade de Informática, Campus de Elviña s/n. A Coruña. Spain. C.P. 15071.

manuel.antonio.novo.perez@udc.es, m.vaamonde@fuac.udc.es

[2]Instituto Tecnolóxico de Matemática Industrial (ITMATI), Edificio Instituto Investigacións Tecnolóxicas, planta -1. Rúa de Constantino Candeira s/n. Campus Vida. Santiago de Compostela. Spain. C.P.15782

marta.rodriguez.barreiro@usc.es

[3]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Facultade de Matemáticas, Rúa Lope Gómez de Marzoa s/n. Campus Vida. Santiago de Compostela. Spain. C.P. 15782.

mariajose.ginzo@usc.es

**RESUMEN**

La monitorización de los medios de extinción durante su trabajo en un incendio es clave para el análisis posterior y la mejora en las tareas de extinción. Se propone un método capaz de identificar las norias activas que siguen las aeronaves durante los trabajos de extinción del incendio, a partir de las posiciones de las mismas. El algoritmo es capaz, a partir de las posiciones aisladas de las aeronaves, de detectar los puntos en los que éstas cargan agua, obtener las trayectorias elípticas que siguen entre los puntos de carga y descarga de agua, y, mediante técnicas de profundidad modal, obtener la trayectoria más representativa, que será identificada como la trayectoria principal, o noria, que siguen las aeronaves en ese instante. Además, se proporciona información de todas las aeronaves asignadas a una noria, y de sus capacidades de descarga agua. En este trabajo se presenta la metodología empleada para obtener las norias activas en cada instante del incendio, y al final se muestran algunos resultados del algoritmo aplicados a incendios reales.

**Palabras y frases clave**: Incendios forestales, aeronaves de extinción, norias.

## 1. INTRODUCCIÓN

La innovación en el sector de los servicios de emergencia es un campo de importancia vital para la sociedad, y en el que la aplicación de las nuevas tecnologías genera un gran potencial. Uno de los elementos más importantes de la transformación tecnológica en la actualidad es el internet de las cosas, que se refiere a tener los objetos conectados a internet, representados mediante datos y metadatos en sistemas informáticos. La aplicación de este nuevo paradigma en el escenario de las aeronaves de extinción permite analizar gran cantidad de información acerca de sus vuelos, como su posición en tiempo real.

Durante su participación en el incendio, las aeronaves realizan circuitos cerrados (habitualmente elípticos) que van desde el punto de carga de agua, hasta el incendio, en dónde descargan. Estos circuitos se denominan norias y su identificación es el objetivo principal del trabajo descrito. Los puntos de carga son conocidos para los operarios, así como la localización aproximada del

incendio. La noria no está fijada de partida, por lo que es el primer piloto en llegar el que decide qué recorrido realizar y en qué puntos cargar y descargar agua. Las trayectorias sucesivas de ésta y otras aeronaves (no necesariamente todas, ya que pueden existir diferentes norias de forma simultánea) siguen ese mismo circuito, hasta que se produce un cambio de noria. Estos cambios de noria son habituales pues el trabajo de extinción de incendios debe adaptarse a la evolución del propio incendio y a las aeronaves disponibles. El conocimiento de estos circuitos puede ser muy útil para mejorar la coordinación de los esfuerzos de extinción de incendios, sobre todo en días en que haya múltiples incendios y/o de gran tamaño, ya que permite mejorar la asignación de las aeronaves a las distintas norias y conocer las zonas del incendio que se están atacando.

Las norias no se registran digitalmente en los registros de vuelo de las aeronaves de extinción, por lo que es necesario determinarlas a partir de los datos de los que se dispone, siendo el objetivo obtener una curva cerrada, ligada a una zona de carga y asociada a las aeronaves que la estén siguiendo. Esto tiene varios problemas asociados, ya que es necesario identificar primero la zona de carga de agua (que tampoco se recoge en los datos de vuelo) y hay que filtrar posiciones en las que las aeronaves no están realizando ninguna noria (por ejemplo, cuando están desplegando brigadistas). A lo anterior cabe añadir que las trayectorias no son perfectamente elípticas y hay que tener en cuenta que cuando están presentes varias aeronaves, pueden seguir o no la misma noria.

Todas estas cuestiones son claves a la hora de desarrollar un algoritmo que permita detectar las norias de un incendio, al que denominaremos "Algoritmo de detección automática de norias". En la Sección 2 se mostrará dicho algoritmo, el cual fue desarrollado en el marco del proyecto Civil UAVs Initiative, financiado por la Xunta de Galicia. Este proyecto busca atraer inversiones en el sector aeroespacial y crear un ecosistema de innovación en el ámbito de la industria de sistemas y vehículos no tripulados. Su objetivo es crear soluciones y productos innovadores para mejorar la prestación de los servicios públicos, haciéndolos más modernos y eficientes.

## 2. ALGORITMO DE DETECCIÓN AUTOMÁTICA DE NORIAS

Como se comentó en la Sección 1, el propósito de este algoritmo es identificar las distintas norias que están activas durante un incendio, indicando las zonas de carga y descarga, así como las aeronaves que participan en ellas.

El algoritmo utiliza las posiciones de las aeronaves, que se registran de forma regular cada pocos segundos. Para cada posición, se cuenta con las coordenadas, la velocidad, la matrícula de la aeronave, la hora a la que se registró y algunos eventos especiales como pueden ser las descargas de agua.

Volviendo a la metodología del algoritmo, en primer lugar se divide la actuación en el incendio en periodos de varios minutos de duración, en los cuales se calcularan las distintas norias durante ese periodo. Este planteamiento permite capturar los distintos cambios tanto en la forma de las norias como en las aeronaves que participan en ellas. La duración de los periodos se puede modificar según el tipo de aeronave, intentando poder contar con datos suficientes para el cálculo de las norias fiables y sin ser demasiado extensos como para que los cambios en las norias distorsionen la estimación. Teniendo en cuenta lo anterior, en cada periodo el algoritmo sigue los siguientes pasos:

1. **Detección de las posiciones de carga de agua para cada aeronave**. Para esto se emplea el algoritmo Density-based spatial clustering of applications with noise (DBSCAN) (Ester et. al 1996), que permite agrupar los puntos próximos identificándolos como miembros de una misma clase. Las variables que se tuvieron en cuenta son las coordenadas de las posiciones, la altitud respecto al nivel del mar y la velocidad de la aeronave, estandarizandolas para no tener problemas con la escala de las variables. Se consideran como cargas aquellos clusters con una velocidad media relativamente baja. En ocasiones, debido a que los datos son bastante irregulares, es necesario utilizar algún filtro más, para eliminar falsas zonas de carga, como puede pasar por ejemplo con algunas paradas aisladas de las aeronaves o si se detecta demasiado cerca de una zona de descarga.

2. **Identificación y suavizado de las trayectorias**. Se clasifican como cargas todas las posiciones cercanas a las clasificadas en el paso anterior, y se eliminan las posiciones consecutivas. Luego se define cada trayectoria como los puntos entre las cargas obtenidas, descartando

aquellas que pasan por varias zonas de carga y que no realizan exactamente una descarga, ya que se identifican con una carga mal identificada en el caso de no haber por lo menos una y con una carga mal identificada en caso de mostrar varias descargas. Luego de esto se transforman los puntos que forman cada trayectoria en representaciones continuas de los vuelos utilizando B-Splines periódicos (De Boor 1978), para garantizar que las norias sean cerradas.

3. **Estimación de norias**. Se comparan las trayectorias de cada aeronave y punto de carga dos a dos, comparando las distancias Hausdorff de todas las trayectorias al resto de trayectoria, diferenciando si se tratan del mismo grupo o de grupos distintos. Luego se aplica un test de Wilcoxon-Mann-Whitney (Bauer 1971) sobre las distancias de ambas clases, de forma que si no difieren significativamente, se asignan todas a la misma noria. Una vez se han agrupado todas las trayectorias, se escoge como representante de cada grupo a la curva con una mayor profundidad modal (Zuo and Serfling 2000), considerando a esta la noria teórica que siguen las aeronaves.

El algoritmo proporciona como salida principal varios objetos ESRI SHAPEFILE tipo línea, representando las norias activas en cada periodo, con información de que aeronaves participan en ellas y las trayectorias que siguieron las aeronaves. También se obtienen varios objetos ESRI SHAPEFILE tipo polígono representando las zonas de carga y descarga, siendo, respectivamente, las elipses de mínima área que contienen los puntos clasificados como cargas y descargas.

El algoritmo está íntegramente programado en el software estadístico R (R Core Team 2021) y es capaz de funcionar de forma automática, estableciendo ciertos parámetros por defecto que varían según las aeronaves con las que va a trabajar, y guardando sus resultados en una base de datos. Esto permite que lo utilicen usuarios que no necesariamente conocen cómo se realizan las estimaciones y además permite analizar la labor realizada en los incendios ya que todos los resultados se pueden recuperar de la base de datos.

## 3. APLICACIÓN A DATOS REALES

En esta sección se muestra un ejemplo de aplicación del algoritmo para un incendio real ocurrido en Santiago de Compostela el 29 de junio de 2020, iniciado a última hora de la tarde en una zona próxima al Complejo Hospitalario Universitario de la Universidad de Santiago de Compostela. El algoritmo se ejecutó utilizando los datos de una única aeronave (un helicóptero) entre las 18 : 15h y las 18.45h.

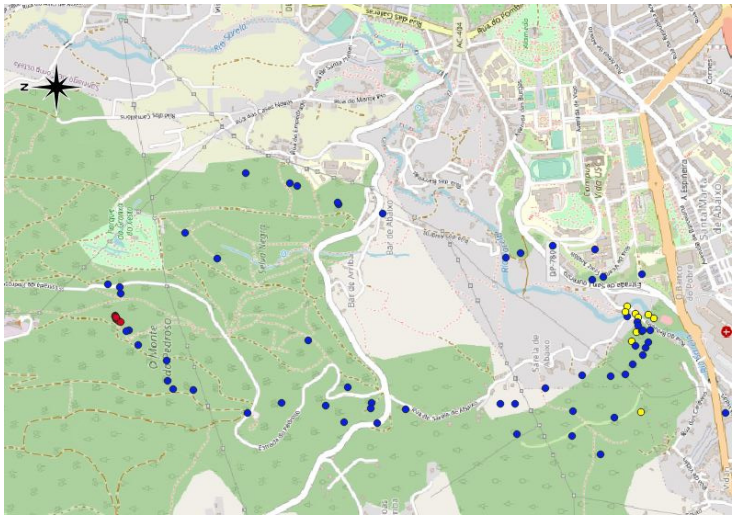


Figura 1: Posiciones registradas de la aeronave, junto con sus descargas (amarillo) y cargas identificadas (rojo).

En la Figura 1 pueden verse los resultados de ejecutar el primer paso del algoritmo, indicando en amarillo las posiciones correspondientes a las descargas de agua y en rojo (en la zona norte) los puntos identificados como cargas y en azul el resto de posiciones.

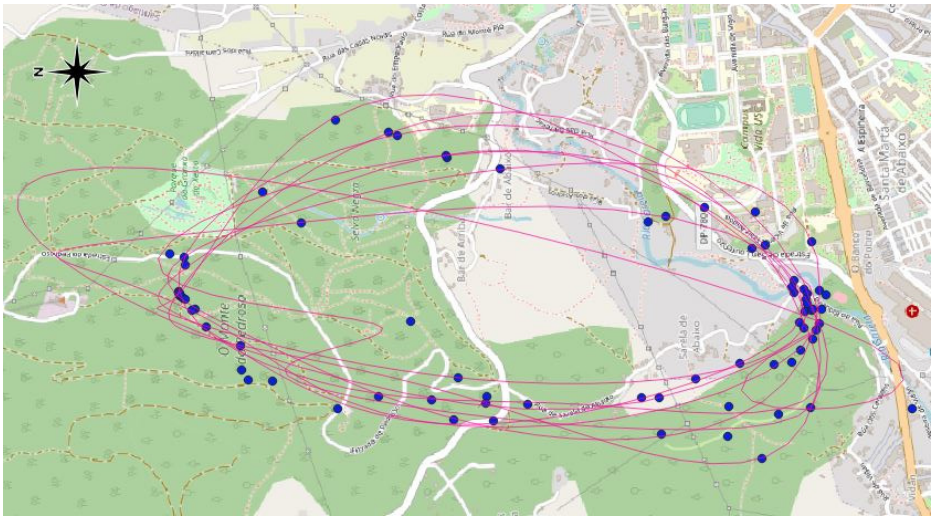


Figura 2: Trayectorias de la aeronave junto con sus posiciones.

En la Figura 2 pueden verse los resultados del segundo paso del algoritmo, que se corresponden con las curvas rosas. Estos son los ciclos carga-descarga-carga estimados por el algoritmo. Se observan algunas trayectorias extrañas debidas al proceso de suavizado, que en este caso no afectan a la estimación de la noria, puesto que se tratan de casos aislados.

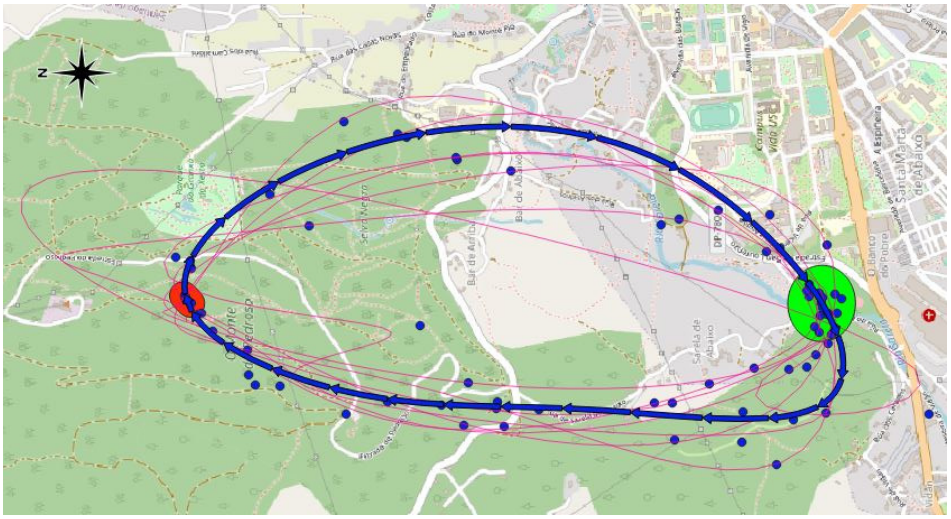


Figura 3: Estimación de la noria (flechas azules) y de las zonas de carga (rojo) y descarga (verde)

Finalmente, en la Figura 3 se pueden ver los resultados de la ejecución del último paso, que se corresponde a los datos de salida del algoritmo. En ella, se puede ver la noria teórica representada por flechas azules (en su sentido de giro) y las zonas de carga (rojo) y de descarga (verde) estimadas por la elipse de mínima área que contiene a los puntos. En este caso se observa que la noria es muy similar a una elipse, lo que concuerda con la concepción teórica que se tiene de ellas.

## 4. CONCLUSIONES

Tras probar el algoritmo en una gran variedad de incendios, se concluye que la solución aportada es bastante robusta y permite estimar tanto norias, como zonas de carga y descarga con cierta flexibilidad y precisión de forma automática. El algoritmo proporciona información muy útil para la coordinación de los incendios, permitiendo por ejemplo tomar decisiones como donde establecer las norias o que aeronaves deben participar en ellas.

De cara a mejorar el algoritmo, las dos principales vías pasarían por mejorar los datos de entrada y registrar las cargas de agua de las aeronaves. Por un lado, aumentar la frecuencia de los datos de posiciones de las aeronaves facilitaría la estimación de sus trayectorias, y éstas serían más precisas. Actualmente se pueden proporcionar trayectorias con formas menos realistas cuando se cuenta con pocos datos para estimarlas, aunque rara vez tienen demasiada influencia en el resultado final. El registro de las cargas de agua (de forma similar que las descargas) simplificaría en gran medida el algoritmo y mejoraría sus resultados.

## REFERENCIAS

David F. Bauer (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, Vol. 67, 687-690.

De Boor, C. (1978). *A practical guide to splines*, Vol. 27, p. 325. New York: springer-verlag.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, Vol. 96, 34, 226-231.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Zuo, Y., & Serfling, R. (2000). General Notions of Statistical Depth Function. *The Annals of Statistics*,28, 461-482. URL: http://www.jstor.org/stable/2674037

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ÍNDICE DE RISCO DE OCORRENCIA DE INCENDIOS EN ESPAÑA

Marta Rodríguez Barreiro[1], Manuel Antonio Novo Pérez [2], Manuel Vaamonde Rivas[2] e María José Ginzo Villamayor[3]

[1]Instituto Tecnolóxico de Matemática Industrial (ITMATI), Edificio Instituto Investigacións Tecnolóxicas, planta -1. Rúa de Constantino Candeira s/n. Campus Vida. Santiago de Compostela. Spain. C.P.15782.
marta.rodriguez.barreiro@usc.es

[2]Universidade da Coruña, Facultade de Informática, Campus de Elviña s/n. A Coruña. Spain. C.P. 15071.
manuel.antonio.novo.perez@udc.es, m.vaamonde@fuac.udc.es

[3]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Facultade de Matemáticas, Rúa Lope Gómez de Marzoa s/n. Campus Vida. Santiago de Compostela. Spain. C.P. 15782.
mariajose.ginzo@usc.es

## RESUMO

Coñecer o risco de que se produza un incendio nunha zona e un momento concretos é fundamental para planificar a xestión dos operativos de extinción. Porén, o risco de ocorrencia de incendios varía en función das características concretas de cada punto da xeografía. Neste traballo proponse a creación dun índice que se adapta ás diferentes áreas do territorio español. Créase a partir dun índice existente desenvolto polo extinto Instituto para a Conservación da Natureza (ICONA), que depende de variables meteorolóxicas e orográficas. A este índice aplícaselle unha modificación baseada na incidencia de incendios ocorridos na zona nos últimos anos, a partires dunha suma ponderada que permite axustar os pesos en función da zona na que se aplique. Unha vez creado o índice, lévase a cabo un proceso de validación que permite ademais elixir os pesos que mellor se adaptan a cada zona. Para isto aplícase un *Support Vector Machine* (SVM) como método de clasificación binaria. Como resultados, amósanse algúns mapas de risco obtidos en Galicia co índice proposto.

**Palabras e frases chave**: incendio forestal, índice de risco, recurrencia, ICONA.

## 1. INTRODUCIÓN

A lacra dos incendios forestais é un dos grandes problemas que afronta o ser humano na actualidade. Segundo o Instituto Galego de Estatística (IGE) en Galicia producíronse máis de 10,000 incendios entre os anos 2015 e 2019. Debido a isto, cada vez son máis os esforzos que se están poñendo nas tarefas de prevención, planificación e extinción de incendios.

Coñecer o risco de que se produza un incendio nunha zona e nun momento concretos é fundamental para optimizar a planificación das actuacións de extinción. É por isto que na literatura existen un gran número de índices, cada un deles creados para uns territorios cunhas características concretas. En xeral, un índice debe ser utilizado únicamente no lugar para o que foi creado, xa que, baixo as mesmas condicións, diferentes zonas poden ter distinto risco de que se produza un incendio. Porén, zonas con condicións similares poden utilizar os mesmos índices, aínda que non fosen creados específicamente para esa área.

Hai diferentes sistemas de índices de risco de incendios. Os máis simples utilizan únicamente variables climatolóxicas, mentres que os máis complexos utilizan un sistema de índices considerando

tamén factores do terreo (Dentoni e Muñoz, 2012). Ademais os diferentes índices sinalan aspectos distintos, algúns reflexan o risco de que se inicie un incendio, mentres que outros fan incidencia tamén na gravidade do mesmo ou a facilidade co que pode propagarse.

En España existen diferentes índices que se aplican nos distintos lugares do territorio. Un dos máis ampliamente utilizados nos equipos de extinción e prevención é o desenrolado polo Instituto para a Conservación da Natureza (ICONA). Debido a isto, este índice será a base do índice proposto neste traballo.

## 2. ÍNDICE DE RISCO DESENVOLTO

O índice de risco desenvolto consiste nunha suma ponderada entre dous índices, un xa existente desenvolto polo extinto ICONA, que a partir de agora será denominado como índice ICONA; e un índice calculado a partir da recurrencia de incendios nos últimos anos na zona, baseado no proposto no plan contra incendios forestais das Illas Baleares (Govern de les Illes Balears, 2015).

### 2.1 Índice ICONA

O índice ICONA é un sistema composto de dous índices, un que representa o risco de que se inicie un incendio, e outro que representa a probabilidade de que éste se propague rapidamente. Como a finalidade do índice a desenvolver é obter o risco de que se inicie un incendio sen importar o seu tamaño, só se utiliza a primeira parte do índice ICONA, a que representa o risco de que se inice un incendio nunha zona e un momento determinados.

O índice ICONA reflexa o risco de que se inicie un incendio a partir da humidade do combustible fino morto, canta menos humidade ten este combustible, máis risco hai de que se inicie un incendio na zona. Para calcular esta humidade, utilízanse unhas táboas de cálculo rápido que, a partir da temperatura ambiente e a humidade relativa, permite obter a humidade do combustible fino morto. Estas táboas poden atoparse no documento do Ministerio de Medio Ambiente que se referencia na bibliografía.

Unha vez que se ten a humidade do combustible fino morto, hai que aplicarlle unha corrección en función da pendente do terreo, a orientación, o sombreado, e a hora e o mes para o que se está calculando o índice. Como no caso anterior, esto faise a través dunhas táboas de cálculo rápido, existen 3 táboas diferentes en función do mes para o que se está a executar o índice.

Unha vez se teñen os dous valores hai que sumalos, e o resultado será o índice de risco ICONA, que variará nunha escala entre 1 e 18, indicando maior risco cos valores máis baixos.

### 2.2 Índice baseado na recurrencia de incendios

Para calcular o índice baseado na recurrencia de incendios, debe dividirse a superficie na que se quere calcular o mapa nunha grella de 10km de resolución. Deste xeito, cóntanse os incendios ocorridos nos últimos cinco anos en cada cela da grella. A continuación, calcúlanse os 17 cuantís do rango do número de incendios, omitindo as celas nas que non ocorreu ningún incendio. A cada cela da grella, asígnaselle un valor, entre 1 e 17, en función do cuantil ó que pertence, dándolle o valor máis baixo ó cuantil máis alto.

Este valor representará o risco de que se inicie un incendio en función da recurrencia de incendios. Pero tamén se quere ter en conta a peligrosidade dos mesmos. Para isto, desenvólvese un procedemento análogo ó anterior, pero esta vez tendo en conta o tipo de chan no que se produxo o incendio: superficie arbolada (Sfa), superficie non arbolada (Sfna) e superficie non forestal (Snf).

Nesta ocasión cóntase o número de incendios ocorridos en cada cela da grella nos últimos 5 anos, para cada tipo de chan. A continuación faise unha suma ponderada das 3 cantidades, para calcular o valor da gravidade dos incendios de acordo a seguinte fórmula:

$$G = \frac{1.5Sfa + 1.25Sfna + Snf}{Sfa + Sfna + Snf}.$$

De forma análoga ó procedemento anterior, calcúlanse os 17 cuantís do rango de valores obtidos da gravidade ($G$), e asígnase un valor entre 1 e 17 a cada cela en función do cuantil ó que pertence, dándolle o valor máis baixo ó cuantil máis alto.

O valor obtido representa a gravidade dos incendios ocorridos nos últimos 5 anos.

Para obter o valor do índice baseado na recurrencia de incendios, o último paso consiste en asignar un valor entre 1 e 17 a cada cela da grella, en función da suma dos dous subíndices calculados.

### 2.3 Índice de risco final

O índice proposto consiste na suma ponderada dos dous índices previamente descritos, como se mostra na seguinte ecuación:

$$ind = w_1 ind_1 + w_2 ind_2,$$

onde $ind_1$ e $ind_2$ son o índice ICONA e o índice baseado na recurrencia de incendios, respectivamente, e $w_1$ e $w_2$ son os pesos asignados.

Os pesos asignados a cada índice variarán dependendo da zona para que éste se calcule segundo ás súas características particulares. Por exemplo, en Galicia, na orixe dos incendios é moi importante o factor antrópico, xa sexa en incendios provocados ou accidentais, debido ó cal existen zonas concretas cunha alta recurrencia de incendios, tendo un incendio ou máis ó ano (Diego *et al.*, 2020).

Nas zonas cunha alta influencia do factor antrópico, o peso asignado ó índice baseado na recurrencia de incendios será máis alto, xa que é este índice o que permite capturar a influencia humana na orixe dos incendios. O proceso de escoller os pesos asignados a cada índice, levouse a cabo no proceso de validación do propio índice.

### 2.4 Validación

A construcción dun novo índice de risco de ocorrencia de incendios non se pode dar por finalizada mentres non se someta a un proceso exhaustivo de validación que demostre a relación existente entre o valor do índice e a existencia do incendio (de Vicente, 2012).

O proceso de validación levado a cabo consiste en dúas partes ben diferenciadas: en primeiro lugar realízase unha análise para comprobar que os valores do índice no momento e na zona nas que hai un incendio son menores (indicando maior risco) que cando non o hai; e compróbase que estas diferenzas son significativas. En segundo lugar, aplícase un SVM para estudiar a bondade das prediccióńs de incendio a partir dos valores do índice.

O proceso de validación dun índice de risco de incendios é complicado debido, entre outras cousas, a que as zonas nas que non hai ningún incendio sempre é moito maior ás zonas nas que si. Ademais, que exista un alto risco de que ocorra un incendio non quere decir que éste chegue a iniciarse necesariamente. Por isto, neste traballo optouse por comparar os resultados obtidos do índice proposto cos do índice ICONA, xa que este último é un índice ampliamente utilizado en España dende a súa creación en 1987 (de Vicente, 2012; Sánchez *et al.*, 2018). No proceso de validación búscase probar que os resultados do índice proposto melloran ós que se obteñen ó aplicar únicamente o índice ICONA.

Para a mostra utilizada, escóllense 60 días nos que existe polo menos un incendio en Galicia, cunha superficie queimada superior a 6ha entre os anos 2015 e 2019.

A combinación de pesos escollida é $w_1 = 1$ e $w_2 = 0$; $w_1 = 0.85$ e $w_2 = 0.15$; $w_1 = 0.75$ e $w_2 = 0.25$; $w_1 = 0.5$ e $w_2 = 0.5$; $w_1 = 0.25$ e $w_2 = 0.75$; onde $w_1$ corresponde ó peso asignado ó índice ICONA e $w_2$ ó peso asignado ó índice baseado na recurrencia.

O proceso de validación consiste en dividir Galicia nunha grella de 10km de resolución, e para cada cela, extraer o valor de cada unha das combinacións de índices propostas. No caso de que nunha cela haxa varios valores (xa que o índice está calculado nun mapa cunha resolución de 200m) escóllese o menor de todos eles. Repítese este proceso para cada unha das 60 datas seleccionadas. Despois, analízase para cada data en qué celas houbo incendio, e compáranse os valores dos índices cando hai incendio e cando non.

Para comprobar se existen diferenzas significativas, aplícanse o T-test e o Mann-Whitney-Wilcoxon test (Vélez e García, 1993).

O seguinte paso do proceso de validación é utilizar un SVM como método de clasificación binaria (Burbidge e Buxton, 2001). Utilízase a mesma mostra de 60 días e a mesma grella de 10km de resolución. A variable resposta do modelo é a existencia ou non de incendio en cada cela, e a variable explicativa o valor do índice de risco calculado para esa cela.

Existen moitas máis celas nas que non existe incendio que nas que si, polo que ó entrenar o modelo o algoritmo clasificará todas as entradas como *Non Incendio* sexa cal sexa a variable explicativa. Para evitar isto, aplícanse uns pesos que permiten balancear as clases segundo o tamaño mostral.

Entrénase un SVM para cada unha das combinacións de pesos escollidas utilizando un 80% da mostra seleccionada, e utilízase o 20% restante para analizar a bondade da clasificación. É esperable que se clasifiquen como *Incendio* moitos casos nos que realmente non existiron, xa que un alto risco de incendios non significa que este ocorra. Porén, o que se quere analizar é que o índice non claifique moitos *Non Incendio* como *Incendio*, xa que esto significaría que o índice sempre representa un alto risco de incendios, e isto na práctica non aporta ningunha información.

Para analizar os resultados obtidos estúdianse as matrices de confusión e aplícase o F-Score (Skolova *et al.*, 2006), que permite resumir a precisón e a sensibilidade do modelo. O F-Score, ademais, é moi útil cando a distribución das clases é desbalanceada e interesa máis non errar nun tipo de clasificación que outro.

Por último, unha vez escollida a mellor combinación de pesos para Galicia, compróbase a eficacia do índice nos incendios ocorridos durante os meses de xullo e agosto no ano 2020 nesta comunidade. Para isto, na grella de 10km de resolución, calcúlase o valor do índice proposto para cada día de xullo e agosto de 2020. Como no caso anterior, cando hai varios valores nunha cela escóllese o menor de todos eles. A continuación, categorízase o índice en 5 niveis: extremo, alto, medio, baixo, moi baixo. Para cada un dos incendios ocorridos (en total 75) estúdiase o nivel de risco no momento no que se produce o incendio.

### 2.5 Resultados

O índice proposto integrouse nun algoritmo programado en R (R Core Team, 2019), que permite calcular o índice sobre un mapa definido polo usuario en 6 momentos do día (de acordo coas táboas do ICONA), podendo analizar a evolución do risco ó largo do día. Os datos de entrada do algoritmo son un modelo dixital do terreo (MDT) que definirá a resolución do mapa final e tamén o mapa sobre o que se calculará o índice, un mapa co sombreado do terreo (como o que proporciona o Sistema de Información sobre Ocupación del Suelo de España, SIOSE), información sobre os incendios ocorridos na zona, e os pesos que se asignan a cada un dos índices que conforman o índice final. Este último parámetro de entrada permite que un mesmo algoritmo se poida adaptar ás diferentes áreas xeográficas, modificando os pesos segundo corresponda.

A continuación móstrase, na Figura 1, un mapa de risco obtido en Galicia o 13-08-2020 ás 14:00 horas. Na parte superior esquerda, pode verse a escala de cores utilizada. Os valores vermellos indican un maior risco, mentres que os azuis indican o risco menor. Pode verse en xeral un risco medio de incendios, e un risco alto na zona sur de Ourense e tamén nalgún punto da costa Oeste. Na zona norte de Galicia acádase o menor risco.

## 3. CONCLUSIÓNS

O índice proposto é un índice capaz de adaptarse ás diferentes zonas de España tendo en conta as súas casuísticas. Ademais, o algoritmo desenvolto permite calcular este índice dun xeito rápido, cunha gran resolución, e sen a necesidade de introducir moitos datos de entrada.

O índice foi sometido a un proceso de validación que permite concluir que os resultados acadados melloran ós resultados que se obteñen utilizando únicamente o índice ICONA, que deixa fóra o factor antrópico que tan importante é nalgunhas zonas, como por exemplo en Galicia.

O feito de que os pesos asignados a cada índice sexan un parámetro de entrada do algoritmo, permite que dun xeito moi sinxelo un mesmo algoritmo se adapte ás diferentes rexións, permitindo incluso calcular o índice ICONA en caso de ser necesario.

Como futuras evolucións deste índice, quedaría considerar máis combinacións de pesos que as escollidas, vendo se se poden mellorar os resultados obtidos. Sería interesante calcular os pesos óptimos a asignar a cada rexión. Ademais, poderíase comparar o índice con outro dos índices de risco existentes na literatura, tal e como se fixo co ICONA.

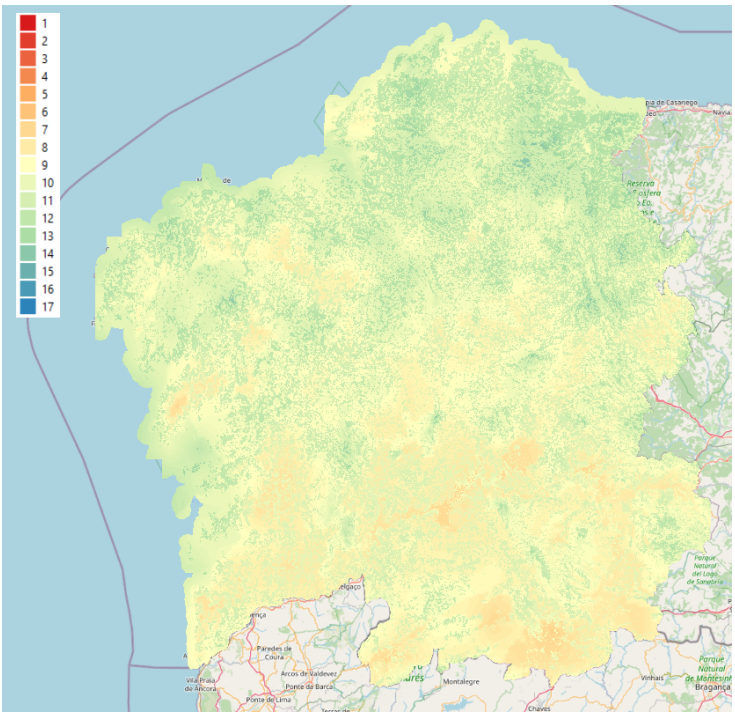



Figura 1: Mapa de risco en Galicia o 13-08-2020 ás 14:00h.

## REFERENCIAS

Burbidge, R., Buxton, B., 2001. An introduction to support vector machines for data mining. Keynote papers, young, OR12-3-15.

Consellería do Medio Rural, 2020. Plan de Prevención y Defensa contra los Incendios Forestlaes de Galicia (PLADIGA). Xunta de Galicia.

Dentoni, M.C., Muñoz, M.M., 2012. Sistemas de Evaluación de Peligros de Incendios. Programa Nacional de Evaluación de Peligro de Incendios y Alerta Temprana. Plan Nacional de Manejo del Fuego. Tech. Rep. 1.

Diego, J., Fernández, M., Rúa, A., 2020. Influencia de la realidad socioeconómica de Galicia en la dinámica de producción de incendios forestales. Boletín de la Asociación de Geógrafos Españoles, 84, 2839, 1-37.

Govern de les Illes Baleares, 2015. IV Plan General de Defensa contra Incendios Forestales de las Islas Baleares. Conselleria d'Agricultura, Medi Ambient i Territori.

Ministerio del Medio Ambiente. Clave Fotográfica para la Identificación de los Modelos de Combustible. Defensa Contra los Incendios Forestales. URL: Clave Fotográfica. Accessed. 05 febrero 2021.

R Core Team, 2019. R: A langage and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL R project.

Sánchez, Y., Martínez-Graña, A., Santos, F., Mateos, M., 2018. Mapping Wildfire Ignition Probability Using Sentinel 2 and LiDAR (Jerte Valley, Cáceres, Spain). Sensors, 18, 826.

Skolova, M., Szpakowicz, S., Japkowicz, N., 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. Conference Paper in Advances in Artificial Intelligence.

Vélez, R., García, A., 1993. Principios de Inferencia Estadística. UNED, España.

de Vicente, F.J., 2012. Diseño de un modelo de riesgo integral de incendios forestales mediante técnicas multicriterio y su automatización en sistemas de información geográfica. Una aplicación en la Comunidad Valenciana. Tesis Doctoral. Universidad Politécnica de Madrid.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# TESTING QUANTILE REGRESSION MODELS WHEN THE RESPONSE VARIABLE IS RIGHT-CENSORED AND THE COVARIATE IS HIGH-DIMENSIONAL

Mercedes Conde-Amboage[1], Ingrid Van Keilegom[2] and Wenceslao González-Manteiga[1]

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

[2] Research Centre for Operations Research and Business Statistics (ORSTAT). Katholieke Universiteit Leuven (KU LEUVEN). Belgium.

## ABSTRACT

Quantile regression was introduced by Koenker and Basset (1978) as a weighted absolute residuals fit which allows to extend some properties of classical least squares estimation to quantile regression estimates. This kind of regression allows a more detailed description of the behaviour of the response variable, adapts to situations under more general conditions of the error distribution (that is, do not require stringent assumptions, such as homoscedasticity or normality) and enjoys properties of robustness. Hereby it facilitates a more complete and robust analysis of the information.

For all that, quantile regression is a very useful statistical technology for a large diversity of disciplines. In particular, quantile regression provides good results when complex data are considered, for instance, when the response variable is right-censored. Along this talk, a new lack-of-fit test for censored quantile regression models with multiple covariates will be presented.

The test is based on the cumulative sum of residuals with respect to unidimensional linear projections of the covariates. The test is then adapting the ideas of Escanciano (2006) to cope with high-dimensional covariates, to the test proposed by Conde-Amboage et al (2021). It will be shown the limit distribution of the empirical process associated with the test statistic. Furthermore, in order to approximate the critical values of the test, a wild bootstrap mechanism is used, which is similar to that proposed by Orbe and Núnez-Antón (2013). In addition, an extensive simulation study and an interesting application to real data will be presented in order to show the behaviour of the new test in practice.

**Keywords**: quantile regression; censored data; lack-of-fit test; bootstrap approach; high-dimensional covariates.

## REFERENCES

Conde-Amboage, M., Van Keilegom, I., and González-Manteiga, W. (2021). A new lack-of-fit test for quantile regression with censored data. Scandinavian Journal of Statistics, 48, 655–688.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. Econometric Theory, 22, 1030–1051.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica, 46, 33–50.

Orbe, J., and Núnez-Antón, V. (2013). Confidence intervals on regression models with censored data. Communications in Statistics-Simulation and Computation, 42, 2140–2159.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# EL PROBLEMA DE LAS DOS MUESTRAS BAJO TRUNCAMIENTO ALEATORIO

Adrián Lago Balseiro[1], Jacobo de Uña Álvarez[1] y Juan Carlos Pardo Fernández[1]

[1]Universidade de Vigo

## RESUMEN

El problema de dos muestras para datos completos ha sido estudiado en profundidad desde hace tiempo en la literatura, siendo el estadístico de Kolmogorov-Smirnov uno de los más empleados para abordarlo. En particular, este estadístico es de distribución libre, lo cual permite su uso inmediato en la práctica. Sin embargo, es muy frecuente la aparición de truncamiento en Análisis de Supervivencia, lo cual provoca que tanto el estimador de la función de supervivencia como el estadístico de Kolmogorov-Smirnov deban ser modificados. Este estimador de máxima verosimilitud de la función de supervivencia es una ligera modificación del propuesto por Kaplan y Meier (1958) para censura aleatoria, véase Klein y Moeschberger (1997). Ahora es posible una extensión del estadístico de Kolmogorov-Smirnov, cuya construcción es similar al caso con datos completos. Al estudiar el comportamiento de este mediante simulaciones se encuentra ya una novedad en el contexto truncado: el estadístico no es de distibución libre, ya que tanto la distribución de la variable de truncamiento como la de tiempo de evento influyen en los cuantiles de la distribución del nuevo estadístico bajo la hipótesis nula. Surgen además dificultades en la definición y en el estudio de este estimador que deben ser tenidas en cuenta, como puede ser la presencia de agujeros (Strzalkowska-Kominak y Stute, 2010).

En este trabajo compararemos esta nueva adaptación del test de Kolmogorov-Smirnov para datos truncados con el comúnmente utilizado test log-rank. Dicho estudio devuelve una primera conclusión que era ya esperable: en una situación de riesgos proporcionales, el log-rank tiene un mejor comportamiento. En el caso de riesgos no proporcionales el comportamiento relativo de ambos tests depende del modelo de truncamiento particular. Se ha comprobado que en el caso de pares concretos de distribuciones para las cuales el test log-rank no tiene potencia, el estadístico de Kolmogorov-Smirnov es capaz de detectar diferencias con potencia razonable. Otra conclusión de nuestro estudio es que el plan de remuestreo bootstrap propuesto para aproximar la distribución nula del estadístico de Kolmogorov-Smirnov lleva a un correcto calibrado del test.

**Palabras y frases clave**: Truncamiento, test de Kolmogorov-Smirnov, test log-rank.

## REFERENCIAS

Kaplan, E.L. y Meier, P. (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457–481.

Klein, J.P. y Moeschberger, M.L. (1997) Survival analysis. Techniques for Censored and Truncated Data. Springer-Verlag.

Strzalkowska-Kominak, E. y Stute, W. (2010) On the probability of holes in truncated samples. Journal of Statistical Planning and Inference, 140, 1519–1518.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# An EM algorithm based estimator for the latency in mixture cure models

A. López-Cheda[1], Y. Peng[2] and M. A. Jácome[1]

[1]Research group MODES, CITIC, Department of Mathematics, University of A Coruña
[2]Queen's Cancer Research Institute, Department of Public Health Sciences and Department of Mathematics and Statistics, Queen's University

## ABSTRACT

Nonparametric methods have attracted much attention in the last few years for cure models. In the literature, to model the effects of covariates on the distribution of the failure time of the susceptible individuals (latency), it is assumed that the cure rate in the model either is a constant or depends on the same covariates as the latency distribution. We propose a new nonparametric estimator for the latency distribution that relaxes the assumption. The estimation, based on the EM algorithm, is readily available for mixture cure models. The finite sample performance of the proposed estimator is assessed in a simulation study. Finally, the proposed method is employed to model the effects of some covariates on the time to bankruptcy among commercial banks insured by the Federal Deposit Insurance Corporation.

**Keywords**: Bootstrap; Censored data; Cure models; EM algorithm; Survival analysis

## 1. INTRODUCTION

We propose a nonparametric method that further extends the work by Xu and Peng (2014) and López-Cheda et al. (2017a, 2017b) to allow different covariates in the cure rate and latency parts. Specifically, the method considers the EM algorithm for fitting the mixture cure model with different covariates in the cure rate and latency parts and demonstrates how it can be used in the proposed nonparametric estimator for the latency survival function. Furthermore, a method to determine the optimal bandwidth in the proposed estimator is presented. The performance of the proposed estimator is assessed in a simulation study. Finally, we apply the proposed estimator to assess the effects of covariates on the time to bankruptcy among commercial banks in the United States in 2006 - 2016.

## REFERENCES

A. López-Cheda, R. Cao, M. A. Jácome and I. Van Keilegom (2017a) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Comput. Stat. Data Anal., 105, 144–165.

A. López-Cheda and M. A. Jácome and R. Cao (2017b) Nonparametric latency estimation for mixture cure models. TEST, 26, 353–376.

J. Xu and Y. Peng (2014) Nonparametric cure rate estimation with covariates. Canadian Journal of Statistics, 42, 1–17.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Sobre la estimación de la distribución bivariante de tiempos de supervivencia sucesivos

Ana Panduro Martín[1], Jacobo de Uña Álvarez[1]

[1]Universidade de Vigo

**RESUMEN**

El modelo de tiempos de eventos sucesivos censurados es de interés en numerosos contextos. En este trabajo se discuten los problemas que aparecen a la hora de proponer estimadores en tal modelo. En particular, se realiza un estudio de simulación para comparar distintos planes de remuestreo bootstrap adaptados a este tipo de datos. Se analiza el modelo cópula y se estudian propiedades importantes del estimador de máxima verosimilitud. Para el caso de una familia paramétrica de cópulas, se revisan los métodos propuestos por Lawless y Yilmaz (2011). Se realizan estudios de simulación que arrojan nuevo conocimiento sobre el comportamiento de los citados métodos y de un plan de remuestreo bootstrap semiparamétrico. Los métodos se aplican a datos médicos reales. Para concluir se introduce el problema de la selección de la cópula.

**Palabras y frases clave**: Análisis de supervivencia, bootstrap, estimación de máxima verosimilitud, funciones cópula, modelos multi-estado.

## 1. INTRODUCCIÓN

El Análisis de Supervivencia multivariante (Hougaard 2000) se ocupa del estudio de dos o más tiempos de vida. Dichos tiempos, en general, van a ser dependientes al estar definidos sobre el mismo individuo. Algunos ejemplos del Análisis de Supervivencia multivariante son los riesgos competitivos, los datos en paralelo y los tiempos de eventos sucesivos. Estos ejemplos también pertenecen al ámbito de los modelos multi-estado (Meira-Machado et al. 2009).

Este trabajo se centra en el estudio de dos tiempos de eventos sucesivos. En este caso, se tiene un par $(X_1, X_2)$ que representa el tiempo hasta el primer evento de interés $(X_1)$ y el tiempo desde el primer evento hasta el segundo $(X_2)$. Estos tiempos sucesivos $(X_1, X_2)$ se encuentran censurados por la derecha por una variable $C$, que se supone independiente del par de tiempos.

Debido a la censura, en vez de los tiempos de interés $X_1$ y $X_2$, se observa la tupla $(T_1, T_2, \delta_1, \delta_2)$, donde $T_1 = \min\{X_1, C\}$, $\delta_1 = \mathbb{I}(X_1 \leq C)$, $T_2 = \min\{X_2, C_2\}$, $\delta_2 = \mathbb{I}(X_2 \leq C_2)$, y donde $C_2 = (C - X_1)\mathbb{I}(X_1 \leq C)$ es la variable de censura del segundo tiempo. Dado que $X_1$ y $X_2$ no se suponen independientes, el segundo tiempo, $X_2$, no va a ser, en general, independiente de la variable que le censura, $C_2$. Esto supone una dificultad a la hora de estimar tanto la distribución conjunta del par de tiempos sucesivos como la marginal del segundo tiempo, para la cual el estimador de Kaplan-Meier será inconsistente en general.

El problema de la estimación de la distribución conjunta de dos tiempos sucesivos censurados ha sido estudiado por numerosos autores, y en este trabajo se ha utilizado como referencia el estimador no paramétrico propuesto por de Uña-Álvarez y Meira-Machado (2008). Siguiendo en el contexto no paramétrico, se presentan dos planes de remuestreo adaptados a este tipo de datos: el bootstrap simple y el obvio. Al contrario de lo que ocurre en el caso univariante, estos métodos no resultan ser equivalentes en el contexto multivariante de este trabajo. Para analizar el comportamiento de dichos métodos bootstrap se realiza un estudio de simulación.

Para modelar los tiempos de eventos sucesivos censurados se puede emplear también el enfoque semiparamétrico propuesto por Lawless y Yilmaz (2011), que utilizan las funciones cópula para modelar la asociación entre los tiempos de interés. Una cópula $\mathcal{C}(u, v)$ es una función de distribución bivariante con distribuciones marginales uniformes en el intervalo $(0, 1)$ (Nelsen 2006). Relacionado con estas funciones, el Teorema de Sklar (1959) establece que cualquier función de distribución bivariante $H(x, y)$ se puede escribir en función de sus marginales y una función cópula. Además, si dichas marginales son continuas, la cópula es única. Las funciones cópula son por este hecho muy útiles a la hora de modelizar la dependencia entre variables aleatorias. En este trabajo son utilizadas para modelizar la asociación de dos tiempos de eventos sucesivos censurados, para los cuales es posible estimar el modelo cópula, es decir, es posible estimar tanto la función cópula como las marginales de los tiempos sucesivos.

En primer lugar, se supone que se conoce completamente la cópula que liga los dos tiempos de interés, y se estudian algunas propiedades de los estimadores de las marginales, como la propiedad de que tan solo tienen masa en los tiempos no censurados. A continuación, se estudia un contexto más general, en el que la cópula no está completamente especificada, sino que se conoce únicamente la familia paramétrica a la que pertenece. Este es el escenario que estudian Lawless y Yilmaz (2011) y que en este trabajo se revisa. Además, en relación con los estimadores propuestos por estos autores, se aportan simulaciones complementarias a las realizadas en dicho artículo. También se analiza el bootstrap semiparamétrico que proponen Lawless y Yilmaz (2011), que es un plan de remuestreo similar al bootstrap obvio nombrado anteriormente. Se han realizado simulaciones que confirman los resultados aportados por dichos autores, y que arrojan además nuevo conocimiento sobre el citado contexto semiparamétrico.

Los métodos revisados en este trabajo, como son la estimación basada en cópulas de las distribuciones marginales de los tiempos sucesivos y el parámetro de la cópula o el bootstrap semiparamétrico, se aplican a un conjunto de datos médicos reales. El trabajo concluye con una introducción al problema de la selección de la cópula. Este problema tiene un gran interés debido al sesgo que aparece en las estimaciones cuando no se especifica bien la familia de cópulas para los tiempos sucesivos. Esto motiva el estudio de un contraste que permita saber si la elección de la familia paramétrica es la correcta. Se propone un posible estadístico para dicho contraste.

## REFERENCIAS

de Uña-Álvarez J, Meira-Machado LF (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters*, 78:2440-2445.

Hougaard P (2000). *Analysis of Multivariate Survival Data*. Springer.

Lawless JF, Yilmaz YE (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53:779-796.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C y Andersen PK (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, 18:195-222.

Nelsen RB (2006). *An Introduction to Copulas*. Springer.

Sklar A (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8:229-231.

*XV Congreso Galego de Estatística e Investigación de Operacións*

*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Automatic selector for the smoothing parameter of Beran's estimator via bootstrap resampling

Rebeca Peláez[1], Ricardo Cao[2], Juan M. Vilar[2]

[1]MODES research group, Department of Mathematics and CITIC, Universidade da Coruña.

[2]MODES research group, Department of Mathematics and CITIC, Universidade da Coruña and ITMATI.

**ABSTRACT**

The generalised product-limit estimator of the conditional survival function depends on a smoothing parameter which is, in practice, unknown. This work proposes a resampling technique to approximate the smoothing parameter of Beran's estimator. It is based on resampling by the smoothed bootstrap and minimising the bootstrap approximation of the mean integrated squared error to find the bootstrap bandwidth. Confidence intervals of the conditional survival curve are also approximated. The behaviour of the method is tested by simulation on several models. Hospitalisation times of COVID-19 patients are analysed to illustrate the performance of the estimator with bootstrap bandwidth.

**Keywords**: Bootstrap; Right censoring; Survival analysis; Conditional survival function; Kernel estimation; Beran's estimator

## 1. INTRODUCTION

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of $(X, Z, \delta)$ with $X$ being the covariate, $Z = \min\{T, C\}$ the observed variable and $\delta = I_{T \leq C}$ the uncensoring indicator. Usually, $T$ is the time until the occurrence of an event and $C$ is the censoring time. The survival function of $T$ is denoted by $S(t)$ and $S(t|x)$ is the conditional survival function of $T$ given $X = x$ evaluated at $t$. The conditional survival function estimator proposed by [1] is given by

$$\widehat{S}_h^B(t|x) = \prod_{i=1}^{n} \left(1 - \frac{I_{\{Z_i \leq t,\ \delta_i = 1\}} w_{n,i}(x)}{1 - \sum_{j=1}^{n} I_{\{Z_j < Z_i\}} w_{n,j}(x)}\right) \tag{1}$$

where

$$w_{n,i}(x) = \frac{K\big((x - X_i)/h\big)}{\sum_{j=1}^{n} K\big((x - X_j)/h\big)}$$

with $i = 1, ..., n$ and $h = h_n$ is the bandwidth for the covariable.

First, finding a method for automatic selection of the smoothing parameter $h$ is interesting. Secondly, the issue of confidence intervals of $S(t|x)$ by means of Beran's estimator is addressed. Bootstrap has become a strong instrument in many statistical applications since it was first introduced by [2]. It is a suitable technique in this context.

## 2. BANDWIDTH SELECTOR FOR BERAN'S ESTIMATOR

In this work, a resampling technique to approximate the smoothing parameters involved in Beran's estimator is defined. Our approach is based on resampling by the smoothed bootstrap and minimising the bootstrap approximation of the mean integrated squared error to find the bootstrap bandwidth. Given $r$ an appropiate pilot bandwidth, the bootstrap resampling algorithm consists of generating $U_i \sim U(0,1)$ and $V_i \sim K$ and obtaining

$$X_i^* = X_{[nU_i]+1} + rV_i,$$

$$Z_i^* = Z_{[nU_i]+1},$$

$$\delta_i^* = \delta_{[nU_i]+1},$$

for each $i = 1, \ldots, n$. The bootstrap sample is formed as $\{(X_i^*, Z_i^*, \delta_i^*)\}_{i=1}^n$.

The optimal smoothing parameter is the bandwidth that minimizes the mean integrated squared error given by:

$$MISE_x(h) = E\Big( \int \big(\widehat{S}_h(t|x) - S(t|x)\big)^2 dt \Big).$$

Then, the bootstrap bandwidth is obtained by minimizing the Monte Carlo approximation of the bootstrap MISE defined as follows

$$MISE_x^*(h) \simeq \frac{1}{B} \sum_{j=1}^{B} \Big( \int \big(\widehat{S}_h^{*(j)}(t|x) - \widehat{S}_r(t|x)\big)^2 dt \Big),$$

where $\widehat{S}_r(t|x)$ is the Beran's survival estimation with pilot bandwidth $r$ using the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$, $\widehat{S}_h^{*(j)}(t|x)$ is the Beran's survival estimation with bandwidth $h$ using the bootstrap resample $\{(X_i^{*(j)}, Z_i^{*(j)}, \delta_i^{*(j)})\}_{i=1}^n$, and $B$ the number of bootstrap resamples.

## 3. SIMULATION STUDY

A simulation study is carried out to analyse the behaviour of the bootstrap algorithm previously described. Several models with different conditional probabilities of censoring were considered. Figure 1 shows the MISE bootstrap functions in two of these scenarios: Model 1 that considers Weibull distribution for life and censoring times and Model 2 that considers exponential life and censoring times. Both models have a conditional probability of censoring equal to 0.5.

A second simulation study focuses on the calculation of confidence intervals of $S(t|x)$ for fixed values of $t$ and $x$. The same resampling technique introduced above and the percentile method are used for this purpose.

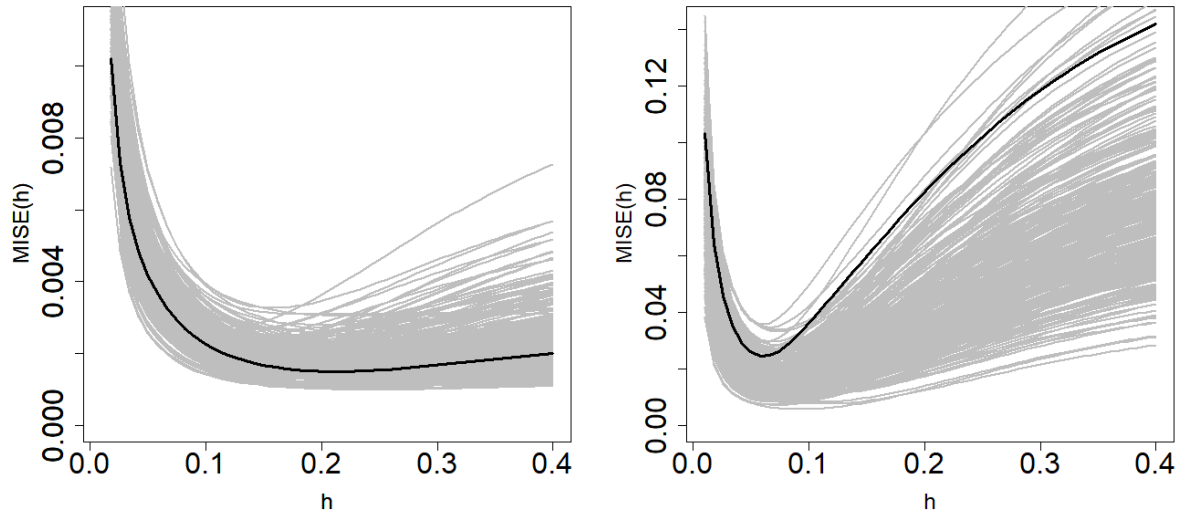



Figure 1: $MISE_x(h)$ function approximated via Monte Carlo using $N = 300$ samples and $MISE^*_x(h)$ approximated via bootstrap using $B = 500$ resamples in Model 1 (left) and Model 2 (right).

Given an appropiate smoothing parameter $h$ and fixed values of time, $t$, and covariate, $x$, the bootstrap confidence interval for a confidence level of $1 - \alpha$ is given by

$$\left(\widehat{S}_h(t|x) - \frac{\rho_{1-\alpha/2}}{\sqrt{nh}},\ \widehat{S}_h(t|x) - \frac{\rho_{\alpha/2}}{\sqrt{nh}}\right),$$

where $\rho_{\alpha/2}$ and $\rho_{1-\alpha/2}$ are the $100\alpha/2$ and $100(1 - \alpha/2)$ percentiles of the resampling distribution of $\sqrt{nh}\big(\widehat{S}^*_h(t|x) - \widehat{S}_r(t|x)\big)$, being $\widehat{S}^*_h(t|x)$ the Beran's survival estimation of the bootstrap resample.

Figure 2 shows the theoretical survival function along with the bootstrap estimation and the bootstrap confidence intervals in one sample from Models 1 and 2 with a conditional probability of censoring equal to 0.5.

## 4. REAL DATA APPLICATION

A brief illustration of the use of the bootstrap technique is provided here. A dataset from the Galician Health Service (SERGAS) with times of hospitalisation and age of 2453 COVID-19 patients in Galicia (Spain) is used. The censoring rate of this dataset is 8.8%. The survival function of the time that COVID-19 patients remain hospitalised in ward is estimated by means of Beran's estimator with bootstrap bandwidth for three different ages. Figure 3 shows Beran's survival estimation with bootstrap bandwidth.

Only 20% of the 40-year-old patients spend more than 15 days in ward. Meanwhile, 40% of COVID-19 positive patient of 60 or 80 years old spend more than 15 days in ward and only 20% of these patients spend more than 25 days in ward.

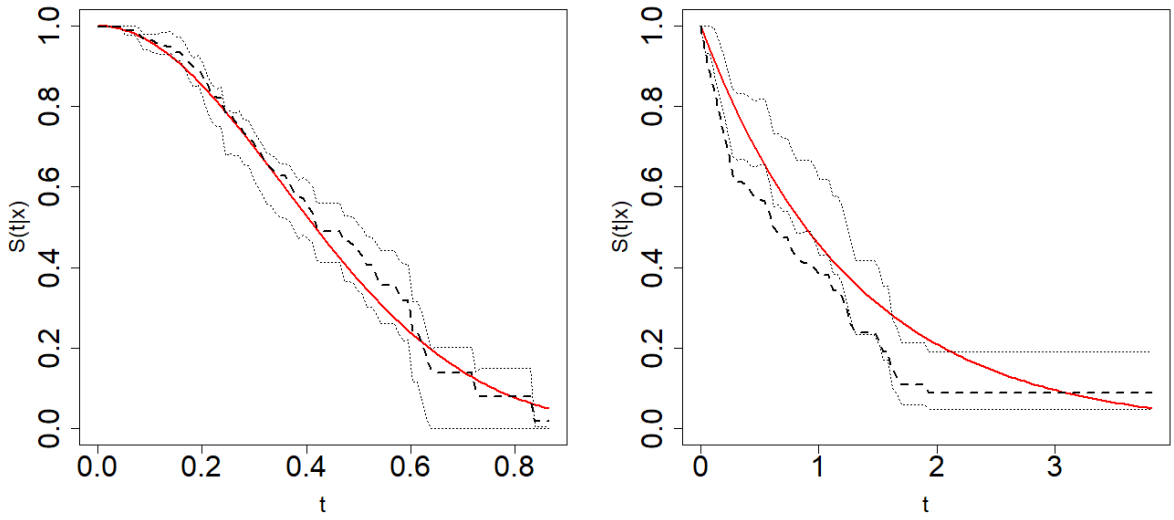



Figure 2: Theoretical survival function (solid line), Beran's estimator with bootstrap bandwidth (dashed line) and the bootstrap confidence intervals (dotted line) for each $t$ in a grid of size $n_t = 100$ in Model 1 (left) and Model 2 (right).

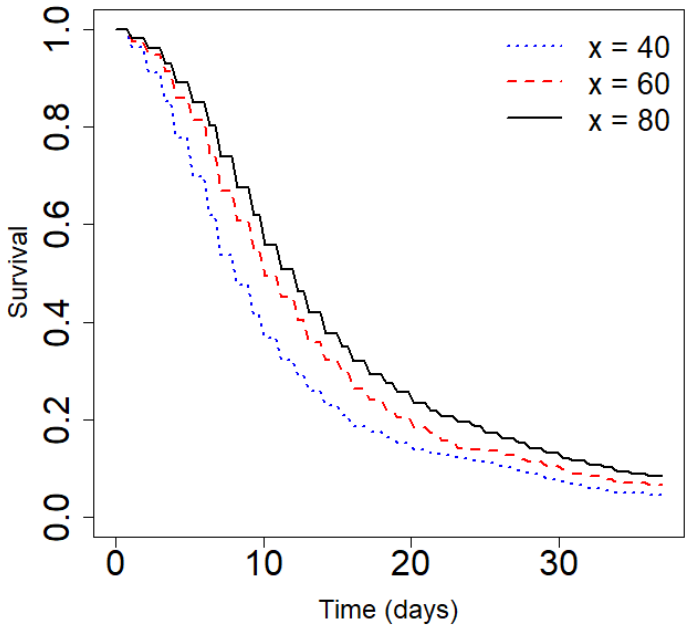



Figure 3: Estimation of $S(t|x)$ for time in ward with Beran's estimator using the optimal bootstrap bandwidth for $x = 40$ (dotted line), $x = 50$ (dashed line) and $x = 80$ (solid line).

### 5. CONCLUSIONS

The results of the simulations show that this bootstrap algorithm provides adequate smoothing parameters to estimate the survival function in this context. The bootstrap bandwidths obtained are similar to the optimal ones and the estimation errors of both are quite similar.

The work of [3] presents a modification of Beran's estimator that involves both a covariate smoothing and a time variable smoothing. We are currently working on an extension of the bootstrap algorithm presented here to select bootstrap bandwidths of the doubly smoothed Beran's estimator, as well as confidence intervals based on it.

## References

[1] Beran, R. (1981) Nonparametric regression with randomly censored survival data. *Technical report*, University of California, Berkeley, 460-463.

[2] Efron, B. (1979) Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, 7, 1-26.

[3] Peláez, R., Cao, Ricardo and Vilar, J.M. (2021) Nonparametric estimation of the conditional survival function with double smoothing, SORT, 42(2), to appear.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Latency function estimation under the mixture cure model when the cure status is available

Wende Clarence Safari 1[1], Ignacio López-de-Ullibarri[1] and María Amalia Jácome[1]

[1]Department of Mathematics, CITIC, Universidade da Coruña

## ABSTRACT

One issue in cure models is to provide estimation and inference of the distribution of the survival time for uncured subjects (latency) while considering the possibility of being cured. Past works have addressed this problem under the assumption that long-term survivors are unidentifiable as cured because of right censoring. However, in some cases this assumption is invalid since some subjects are known to be cured, e.g., when a medical test ascertains that a certain disease has entirely disappeared after treatment. In this paper, we propose a nonparametric latency estimator that extends the available nonparametric estimator in the mixture cure model that ignores the cure status information. Some asymptotic results of the proposed estimator are established, its behavior is analyzed throughout a simulation study, and compared its performance with three other estimators that are available in the literature. Among all the estimators studied, the proposed estimator showed the best behavior. Finally, the proposed estimator is applied to a medical data to study the length of hospital stay of COVID-19 patients requiring intensive care.

**Keywords**: Bootstrap bandwidth; Censoring; Cure model; COVID-19; Nadaraya-Watson weights.

## 1. INTRODUCTION

In survival analysis it is commonly assumed that the considered event will happen when there is a sufficient follow-up time. However, there are many instances where the event of possible interest will not occur. Some cancer patients are known to never relapse or die from cancer, some people are happy enough at their jobs that they never leave, etc. The proportion of those whose event is certain not to occur are considered "statistically cured" (or long-term) survivors and those who will develop the event are known to be "uncured" (or susceptible) subjects. Cure models (Peng and Yu, 2021) have been developed to address this issue. Given the cure model framework, in this work we consider the well-known mixture cure model (MCM), which characterizes, separately, the probability to be cured and the distribution of the time-to-event (latency) for the uncured subjects. Many (semi)parametric and nonparametric estimation methods have been proposed for estimating the latency function in MCM under the assumption that long-term survivors are unidentified as cured because of right censoring (Maller and Zhou, 1996; Patilea and Van Keilegom, 2017; López-Cheda et al., 2017; Amico and Van Keilegom, 2018).

Being able to identify who is cured or not among censored subjects is indeed relevant in several studies. For example, diagnostic procedures in medical studies are available to provide further information on whether a subject has been cured or not (Wu et al., 2014a,b). Also, for some types of cancer it is extremely unlikely to have any recurrence later than a given fixed time after treatment, known as cure threshold (Laska and Meisner, 1992, Taylor, 1995; Nieto-Baraja and Yin, 2008; Bernhardt, 2016). Through extensive simulations, all these studies demonstrated that these extended mixture cure models provide more efficient and less biased estimations than the standard mixture cure models. Here we propose a nonparametric latency estimator in the presence of the cure status information. For the choice of the smoothing parameter, we propose a bandwidth selector based on the bootstrap.

## 2. ESTIMATION WHEN THE CURE STATUS IS PARTIALLY AVAILABLE

### 2.1 MODEL NOTATION

Let $Y$ be the survival time, $\mathbf{X}$ a vector of covariates and $F(t \mid \mathbf{x})$ the distribution function of $Y$ conditional on $\mathbf{X} = \mathbf{x}$. Let $C^*$ be a random censoring time with conditional distribution function $G(t \mid \mathbf{x})$. So, instead of observing $Y$, only $T^* = \min(Y, C^*)$ and $\delta = \mathbf{1}(Y < C^*)$ can be observed. The random variables $Y$ and $C^*$ are assumed to be conditionally independent given $\mathbf{X} = \mathbf{x}$. We set $Y = \infty$ if the subject is cured. Let $\nu = \mathbf{1}(Y = \infty)$ be the indicator of being cured. Note that $\nu$ is partially observed because $\delta = 1$ implies $\nu = 0$, but in the general situation $\nu$ is unknown when $\delta = 0$. When the cure status is partially known, $\nu = 1$ is also observed for some censored individuals.

To accommodate the cure status information, we include an additional random variable $\xi$, which indicates whether the cure status is known ($\xi = 1$) or not ($\xi = 0$). Further, let the censoring distribution be an improper distribution function $G(t \mid \mathbf{x}) = (1 - \pi(\mathbf{x}))\, G_0(t \mid \mathbf{x})$, so with probability $\pi(\mathbf{x})$ the censoring variable is $C^* = \infty$, and with probability $1 - \pi(\mathbf{x})$ the value of the censoring variable $C^*$ corresponds to the value of a random variable $C$ with proper continuous distribution function $G_0(t \mid \mathbf{x})$. A cured individual is identified with probability $P(\xi = 1 \mid \nu = 1, \mathbf{X} = \mathbf{x}) = P(C^* = \infty \mid \mathbf{X} = \mathbf{x}) = \pi(\mathbf{x})$. In this setup, the data actually observed are $\{(\mathbf{X}_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$, where the observed time for the individuals not identified as cured is $T_i = T_i^*$ and if an individual is cured, the corresponding observed time is not $T_i = \infty$ but $T_i = C_i$. Hence, the observations $\{(\mathbf{X}_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$ can be classified into three groups: (a) the individual is observed to have experienced the event and therefore known to be uncured ($\mathbf{X}_i, T_i = Y_i, \delta_i = 1, \xi_i = 1, \xi_i\nu_i = 0$); (b) the lifetime is censored and the cure status is unknown ($\mathbf{X}_i, T_i = C_i, \delta_i = 0, \xi_i = 0, \xi_i\nu_i = 0$); and (c) the lifetime is censored and the individual is known to be cured ($\mathbf{X}_i, T_i = C_i, \delta_i = 0, \xi_i = 1, \xi_i\nu_i = 1$). In standard cure models when the cure status is unknown for the censored observations, only groups (a) and (b) are considered.

The probability of cure is $1 - p(\mathbf{x}) = P(Y = \infty \mid \mathbf{X} = \mathbf{x})$, and the conditional survival function of the uncured individuals, also known as latency, is $S_0(t \mid \mathbf{x}) = P(Y > t \mid Y < \infty, \mathbf{X} = \mathbf{x})$. The mixture cure model specifies the survival function $S(t \mid \mathbf{x}) = P(Y > t \mid \mathbf{X} = \mathbf{x})$ as

$$S(t \mid \mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x}) S_0(t \mid \mathbf{x}). \tag{1}$$

One key issue in cure models is identifiability. Recent works (Hanin and Huang, 2014), and references therein, have discussed the condition needed for the MCM to be identifiable. They demonstrated that $\lim_{t\to\infty} S_0(t \mid \mathbf{x}) = 0$ for all $\mathbf{x}$ is a sufficient condition for identifiability of model (1). In this work, we assume this condition holds.

### 2.2 PROPOSED ESTIMATOR

Without loss of generality, let us consider a continuous covariate $X$ with density function $m(x)$. Owing to the relationship in (1), the latency function can be written in terms of the survival function and cure probability as follows:

$$S_0(t \mid x) = \frac{S(t \mid x) - (1 - p(x))}{p(x)}. \tag{2}$$

Safari et al. (2021a) proposed the generalized product-limit estimator of the conditional survival function $S(t \mid x)$ when the cure status is partially known, which is

$$\widehat{S}_h^c(t \mid x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)\, \mathbf{1}\left(T_{(i)} \le t\right)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x)\, \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)}\right), \tag{3}$$

and the resulting estimator was used to construct an estimator of the cure probability $1 - p(x)$ (Safari et al., 2021b), which is

$$1 - \widehat{p}_h^c(x) = \widehat{S}_h^c\left(T_{(n)}^1 \mid x\right) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x)\, \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)}\right), \tag{4}$$

where $X_{[i]}$, $\delta_{[i]}$, $\xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} \leq \cdots \leq T_{(n)}$, $B_{h[i]}(x)$ are the Nadaraya-Watson (NW) weights,

$$B_{h[i]}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h\left(x - X_j\right)},$$

$K_h(\cdot) = K(\cdot/h)/h$ is a kernel function $K(\cdot)$ rescaled with bandwidth $h$ and $T^1_{(n)}$ is the largest uncensored observed time. Here, in light of (3) and (4) we use the relation in (2) to construct a nonparametric estimator of the latency function when the cure status information is available. Note that the optimal bandwidth for $\widehat{S}^c_h(t \mid x)$ in (3) is not necessarily the optimal bandwidth for $1-\widehat{p}^c_h(x)$ in (4), therefore, we propose a more general estimator that uses two different bandwidths for estimating $S(t \mid x)$ and $1 - p(x)$:

$$\widehat{S}^c_{0,h_1,h_2}(t \mid x) = \begin{cases} \dfrac{\widehat{S}^c_{h_2}(t \mid x) - (1 - \widehat{p}^c_{h_1}(x))}{\widehat{p}^c_{h_1}(x)} & \text{if } \ 0 \leq t \leq T^1_{(n)} \ \text{ and } \ \widehat{S}^c_{h_2}(t \mid x) > 1 - \widehat{p}^c_{h_1}(x) \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In this way, we are guaranteed that the proposed estimator is a proper non-increasing survival function and it goes to 0 as time $t$ goes to infinity. Note that if $h = h_1 = h_2$ then the estimator in (5) reduces to the following estimator:

$$\widehat{S}^c_{0,h}(t \mid x) = \frac{\widehat{S}^c_h(t \mid x) - (1 - \widehat{p}^c_h(x))}{\widehat{p}^c_h(x)}. \tag{6}$$

Although the estimator in (6) guarantees a proper survival function it might not be flexible enough when the optimal bandwidths for $\widehat{S}^c_h(t \mid x)$ and $1 - \widehat{p}^c_h(x)$ are quite different. In the unconditional case, the estimators in (5) and (6) become

$$\widehat{S}^c_{0,n}(t) = \frac{\widehat{S}^c_n(t) - (1 - \widehat{p}^c_n)}{\widehat{p}^c_n}$$

where

$$\widehat{S}^c_n(t) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} \mathbf{1}\left(T_{(i)} \leq t\right)}{n - i + 1 + \sum_{j=1}^{i-1} \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)}\right)$$

is the generalization of the Kaplan-Meier estimator of the survival function to the presence of cured individuals with some of them identified as cured proposed by Safari et al., (2021a), and

$$1 - \widehat{p}^c_n = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n - i + 1 + \sum_{j=1}^{i-1} \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)}\right),$$

is the unconditional estimator of the probability of cure by Safari et al., (2021b).

## 3. APPLICATION TO COVID-19 DATA

We illustrate the performance of the new estimator using the COVID-19 data for patients requiring admission to the ICU during the first wave (March 6 – May 5, 2020) of the pandemic in Galicia. During this period, nearly 2,380 patients were hospitalized for at least one day. The Galician Healthcare Service experienced a significant need of hospital beds, ICU beds, and other health care resources. Thus, an accurate prediction of the total time spent in hospital ward until admission to ICU was crucial for decision making and suitable planning. This analysis aims to illustrate the time from the hospital ward to the ICU admission for patients admitted to the ICU (latency function), given age and sex as covariates of interest. Among hospitalized patients, 197 (8.3%) patients needed ICU care. A total of 1,638 (68.8%) patients were discharged alive before entering ICU, and 328 (13.8%) had died before entering ICU, hence, known to be "cured" from the ICU admission.

## REFERENCES

Amico, M., and Van Keilegom, I. (2018). Cure models in survival analysis. Annual Review of Statistics and Its Application, 5, 311–342.

Bernhardt, P. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. Statistics in Medicine, 25, 4607–4623.

Hanin, L., and Huang, L. S. (2014). Identifiability of cure models revisited. Journal of Multivariate Analysis, 130, 261–274.

Laska, E.M. and Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. Biometrics, 48(4), 1223–1234.

López-Cheda, A. Jácome, M.A. and Cao, R. (2017). Nonparametric latency estimation for mixture cure models. TEST, 2, 353–376.

Maller, R. A., and Zhou, X. (1996). Survival analysis with long-term survivors. New York: Wiley.

Nieto-Barajas, L. E., and Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. Scandinavian Journal of Statistics, 35(3), 540–556.

Patilea, V., and Van Keilegom, I. (2020). A general approach for cure models in survival analysis. The Annals of Statistics, 48(4), 2323–2346.

Peng, Y., and Yu, B. (2021). Cure Models: Methods, Applications, and Implementation. CRC Press.

Safari W.C., López-de-Ullibarri I. and Jácome M. A. (2021a). A product-limit estimator of the conditional survival function when cure status is partially known. Biometrical Journal, 63(5), 984–1005.

Safari W.C., López-de-Ullibarri I. and Jácome M. A. (2021b). A product-limit estimator of the conditional survival function when cure status is partially known. Submitted, available at *https://dm.udc.es/preprint/main_paper_cure_rate_Safari_et_al.pdf*

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models. Biometrics, 899–907.

Wu, Y., Lin, Y., Li, C. S., Lu, S. E., and Shih, W. J. (2014a). Asymptotic efficiency of an exponential cure model when cure information is partially known. International Journal of Statistics and Probability, 3(3), 1–17.

Wu, Y., Lin, Y., Lu, S. E., Li, C. S., and Shih, W. J. (2014b). Extension of a Cox proportional hazards cure model when cure information is partially known. Biostatistics, 15(3), 540–554.

# Premios modalidade A

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

**Bagging cross-validated bandwidth selection in nonparametric regression estimation with applications to large-sized samples**

Daniel Barreiro-Ures[1], Ricardo Cao[1] and Mario Francisco-Fernández[1]

[1]Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

**ABSTRACT**

Cross-validation is a well-known and widely used bandwidth selection method in nonparametric regression estimation. However, this technique has two remarkable drawbacks: (i) the large variability of the selected bandwidths, and (ii) the inability to provide results in a reasonable time for very large sample sizes. To overcome these problems, bagging cross-validation bandwidths are analyzed in this paper. This approach consists in computing the cross-validation bandwidths for a finite number of subsamples and then rescaling the averaged smoothing parameters to the original sample size. Under a random-design regression model, asymptotic expressions up to a second-order for the bias and variance of the leave-one-out cross-validation bandwidth for the Nadaraya–Watson estimator are obtained. Subsequently, the asymptotic bias and variance and the limit distribution for the bagged cross-validation selector are derived. Suitable choices of the number of subsamples and the subsample size lead to an $n^{-1/2}$ rate for the convergence in distribution of the bagging cross-validation selector, outperforming the rate $n^{-3/10}$ of leave-one-out cross-validation. Several simulations and an illustration on a real dataset related to the COVID-19 pandemic show the behavior of our proposal and its better performance, in terms of statistical efficiency and computing time, when compared to leave-one-out cross-validation.

**Keywords**: bagging, cross-validation, Nadaraya–Watson, regression, subsampling

# 1 Introduction

The study of a variable of interest depending on other variable(s) is a common problem that appears in many disciplines. To deal with this issue, an appropriate regression model setting up the possible functional relationship between the variables is usually formulated. As part of this analysis, the unknown regression function, describing the general relationship between the variable of interest and the explanatory variable(s), has to be estimated. This task can be carried out using nonparametric methods that do not assume any parametric form for the regression function, providing flexible procedures and avoiding misspecification problems. Among the available nonparametric approaches, kernel-type regression estimators (Wand and Jones, 1995) are perhaps the most popular. To compute this type of estimators the user has to select a kernel function (typically a density function) and a bandwidth or smoothing parameter that regulates the amount of smoothing to be used, which in turn determines the trade-off between the bias and the variance of the estimator. Although the choice of the kernel function is of secondary importance, the smoothing parameter plays a crucial role. In this regard, numerous contributions have been made over the last decades, providing methods to select the bandwidth. These approaches include, among others, cross-validation methods (Härdle et al., 1988) and plug-in selectors (Ruppert et al., 1995). In Köhler et al. (2014), a complete review and an extensive simulation study of different data-driven bandwidth selectors for kernel regression are presented. Due to their wide applicability and the good performance obtained in this complete comparison, in the present paper, we focus on analyzing cross-validation bandwidth selection techniques.

Cross-validation is a popular method of model selection that precedes an early discussion of the method by Stone (1974). In its simplest form, cross-validation consists of splitting the dataset under study into two parts, using one part to fit one or more models, and then predicting the data in the second part with the models so-built. In this way, by not using the same data to fit and validate the models, it is possible to objectively compare the predictive capacity of different models. The leave-one-out version of cross-validation (of interest in the present paper) is somewhat more involved. It excludes one datum from the dataset, fits a model from the remaining observations, uses this model to predict the datum left out, and then repeats this process for all the data.

The present work studies the leave-one-out cross-validation bandwidth selection method and the application of bagging (Breiman, 1996) to this procedure. We derive some asymptotic properties of the corresponding selectors when considering a random-design regression model and the Nadaraya–Watson kernel-type estimator is used. The Nadaraya–Watson estimator can be seen as a particular case of a wider class of nonparametric estimators, the so-called local polynomial estimators (Stone, 1977; Cleveland, 1979; Fan, 1992), when performing a local constant fit. Given a random sample of size $n$, bagging cross-validation consists of selecting $N$ subsamples of size $r < n$, each without replacement, from the $n$ observations. One then computes a cross-validation bandwidth from each of the $N$ subsets, averages them, and then scales the average down appropriately to account for the fact that $r < n$. It is well-known that the use of bagging can lead to substantial reductions in the variability of an estimator that is nonlinear in the observations (see Friedman and Hall, 2007), as occurs in the case of the cross-validation criterion function. The use of bagging in conjunction with cross-validation for bandwidth selection has already been studied in the case of kernel density estimation by several authors (see, for example Barreiro-Ures et al., 2020; Hall and Robinson, 2009). In addition to the potential improvement in statistical precision, even in the case of small sample sizes, the use of bagging (with appropriate elections of $r$ and $N$) can drastically reduce computation times, especially for very large sample sizes. Note that the complexity of cross-validation is $O(n^2)$, while the complexity of bagging cross-validation is $O(Nr^2)$. Larger reductions in computation time can also be additionally achieved with the application of binning techniques in the bagging procedure.

Apart from the theoretical analysis of the cross-validation bandwidth selection methods, another goal of this study is to apply the techniques studied in the present work to a dataset related to the current COVID-19 pandemic. In particular, using a moderately large sample, provided by the Spanish Center for Coordinating Sanitary Alerts and Emergencies, consisting of the age and the time in hospital of people infected with COVID-19 in Spain, we are interested in studying the relationship between those two variables by means of the Nadaraya–Watson estimator. Apart from its purely epidemiological interest and due to the considerable size of the sample, this dataset is also useful to put into practice the techniques analyzed in the present paper.

## 2 Regression model and Nadaraya–Watson estimator

Let $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an independent and identically distributed (i.i.d.) sample of size $n$ of the two-dimensional random variable $(X, Y)$, drawn from the nonparametric regression model:

$$Y = m(X) + \varepsilon, \tag{1}$$

where $m(x) = \mathrm{E}(Y \mid X = x)$ denotes the regression function, and $\varepsilon$ is the error term, satisfying that $\mathrm{E}(\varepsilon \mid X = x) = 0$ and $\mathrm{E}\left(\varepsilon^2 \mid X = x\right) = \sigma^2(x)$.

The Nadaraya–Watson estimator or local constant estimator (Nadaraya, 1964; Watson, 1964) offers a nonparametric way to estimate the unknown regression function, $m$. It is given by:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}, \tag{2}$$

where $h > 0$ denotes the bandwidth or smoothing parameter and $K$ the kernel function. As pointed out in the introduction, the value of the bandwidth is of great importance since it determines the

amount of smoothing performed by the estimator and, therefore, heavily influences its behavior. Thus, in practice, data-driven bandwidth selection methods are needed.

Optimal bandwidths often refer to smoothing parameter values that mimize some error criterion function. These functions are typically expected loss, in some sense. When the aim is predicting the response variable, $Y$, given the value of the explanatory variable, $X$, it is natural to consider expectations conditionally on the observed explanatory sample, $(X_1, \ldots, X_n)$. However, the focus of this paper is estimating the regression function on its own. Thus an unconditional expected loss view is adopted. Of course, there exist arguments in favor of both type of criteria. More details on this issue can be found in Köhler et al. (2014).

When adopting an unconditional view, a possible way to select a (global) optimal bandwidth for (2) consists in minimizing, for instance, the mean integrated squared error (MISE), a (global) optimality criterion defined as:

$$M_n(h) = \mathrm{E}\left[\int \{\hat{m}_h(x) - m(x)\}^2 f(x)\, dx\right], \tag{3}$$

where $f$ denotes the marginal density function of $X$. The bandwidth that minimizes (3) is called the MISE bandwidth and it will be denoted by $h_{n0}$, that is,

$$h_{n0} = \arg\min_{h>0} M_n(h). \tag{4}$$

The MISE bandwidth depends on $m$ and $f$ and, since in practice both functions are often unknown, $h_{n0}$ cannot be directly calculated. However, it can be estimated, for example, using the cross-validation method.

In the following section, we present the leave-one-out cross-validation bandwidth selection criterion and provide the asymptotic properties of the corresponding selector when using the estimator (2) and considering the regression model (1).

## 3 Cross-validation bandwidth

Cross-validation is a method that offers a criterion for optimality which works as an empirical analogue of the MISE and so it allows us to estimate $h_{n0}$. The cross-validation function is defined as:

$$CV_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) - Y_i\right\}^2, \tag{5}$$

where $\hat{m}_h^{(-i)}$ denotes the Nadaraya–Watson estimator constructed using $\mathcal{X} \setminus \{(X_i, Y_i)\}$, that is, leaving out the $i$-th observation,

$$\hat{m}_h^{(-i)}(x) = \frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{n} K_h\left(x - X_j\right) Y_j}{\sum\limits_{\substack{j=1\\ j\neq i}}^{n} K_h\left(x - X_j\right)}. \tag{6}$$

Hence, the cross-validation bandwidth, $\hat{h}_{CV,n}$, can be defined as

$$\hat{h}_{CV,n} = \arg\min_{h>0} CV_n(h). \tag{7}$$

It is well-known that under suitable regularity conditions, up to first order,

$$M_n(h) = B_1 h^4 + V_1 n^{-1} h^{-1} + O\left(h^6 + n^{-1}h\right),$$

where

$$\begin{aligned} B_1 &= \frac{1}{4}\mu_2(K)^2 \int \left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\}^2 f(x)\, dx, \\ V_1 &= R(K)\int \sigma^2(x)\, dx, \end{aligned}$$

with $R(g) = \int g^2(x)\,dx$ and $\mu_j(g) = \int x^j g(x)\,dx$, $j = 0, 1, \ldots$, provided that these integrals, as well as $B_1$ and $V_1$, exist finite. Then, the first-order term of the MISE bandwidth, $h_n$, has the expression $h_n = C_0 n^{-1/5}$, where

$$C_0 = \left(\frac{V_1}{4B_1}\right)^{1/5}.$$

In order to obtain the asymptotic properties of (7) as an estimator of (4), it is necessary to study certain moments of (5) and its derivatives. However, the fact that the Nadaraya–Watson estimator has a random denominator makes this a very difficult task. To overcome this problem, it will be useful to work with an approximation of $\hat{m}_h(x)$. For this, note that the Nadaraya–Watson estimator can be written as

$$\hat{m}_h(x) = A + B + C + D + E + F, \tag{8}$$

where

$$\begin{array}{ccc} A = \frac{\hat{a}}{e}, & B = \frac{a(e-\hat{e})}{e^2}, & C = \frac{(\hat{a}-a)(e-\hat{e})}{e^2}, \\ D = \frac{a}{e}\frac{(e-\hat{e})^2}{e^2}, & E = \frac{\hat{a}-a}{e}\frac{(e-\hat{e})^2}{e^2}, & F = \frac{\hat{a}}{\hat{e}}\frac{(e-\hat{e})^3}{e^3}, \end{array}$$

with

$$\begin{array}{cc} a = m(x)f(x), & e = f(x), \\ \hat{a} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\,Y_i, & \hat{e} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i). \end{array}$$

Expression (8) splits $\hat{m}_h(x)$ as a sum of five ratios with no random denominator plus an additional term, $F$, which has a random denominator. However, both $E$ and $F$ are negligible with respect to the other terms. Thus, one may consider the modified version of the Nadaraya–Watson estimator given by $\tilde{m}_h(x) = A + B + C + D$, that is:

$$\tilde{m}_h(x) = m(x) + \frac{1}{n^2 f(x)^2}\sum_{j=1}^{n}\sum_{k=1}^{n} K_h(x - X_j)\{Y_j - m(x)\}\{2f(x) - K_h(x - X_k)\}, \tag{9}$$

which can be seen as a quadratic approximation of $\hat{m}_h(x)$, where the terms $E$ and $F$ are omitted due to their "cubic negligibility". In practice, (9) is unobservable and, therefore, it does not define an estimator but a theoretical approximation of (2). This decomposition of $\hat{m}_h(x)$ is in turn inspired by a similar approach proposed in Barbeito (2020). There, a linear approximation of the Nadaraya–Watson estimator was considered and so only the terms $A$ and $B$ were taken into account, leading to the simpler expression

$$\bar{m}_h(x) = m(x) + \frac{1}{nf(x)}\sum_{i=1}^{n} K_h(x - X_i)\{Y_i - m(x)\}. \tag{10}$$

Following this approach, (9) could be used to define a theoretical approximation of the MISE function defined in (3), namely

$$\tilde{M}_n(h) = \int \left[\mathrm{E}\{\tilde{m}_h(x)\} - m(x)\right]^2 f(x)\,dx + \int \mathrm{var}\{\tilde{m}_h(x)\}\,f(x)\,dx.$$

The bandwidth that minimizes $\tilde{M}_n(h)$ will be denoted by $\tilde{h}_{n0}$. On the other hand, (9) can also be used to define a modified version of the cross-validation criterion,

$$\widetilde{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{m}_h^{(-i)}(X_i) - Y_i\right\}^2, \tag{11}$$

where $\tilde{m}_h^{(-i)}$ denotes the leave-one-out version of (9) without the $i$-th observation, that is,

$$\begin{aligned} \tilde{m}_h^{(-i)}(x) \quad = \quad & m(x) + \frac{1}{(n-1)^2 f(x)^2}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i}}^{n} K_h(x - X_j)\{Y_j - m(x)\} \\ & \{2f(x) - K_h(x - X_k)\}. \end{aligned} \tag{12}$$

The bandwidth that minimizes (11) will be denoted by $\tilde{h}_{CV,n}$. Using Taylor expansions, the following approximation can be obtained:

$$\begin{aligned}\tilde{h}_{CV,n} - \tilde{h}_{n0} &\approx -\frac{\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})}{\tilde{M}''_n(\tilde{h}_{n0})} \\ &+ \frac{\left\{\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})\right\}\left\{\widetilde{CV}''_n(\tilde{h}_{n0}) - \tilde{M}''_n(\tilde{h}_{n0})\right\}}{\tilde{M}''_n(\tilde{h}_{n0})^2},\end{aligned} \tag{13}$$

where the second term of (13) is negligible with respect to the first one and is assumed not to contribute to the bias and the variance of $\tilde{h}_{CV,n}$. Since the first-order terms of $\mathrm{E}\left\{\widetilde{CV}^{k)}_n(h)\right\}$ and $\tilde{M}^{k)}_n(h)$ coincide for every $k \geq 1$, we need to calculate the second-order terms of both $\mathrm{E}\left\{\widetilde{CV}'_n(\tilde{h}_{n0})\right\}$ and $\tilde{M}'_n(\tilde{h}_{n0})$ in order to analyze the bias of the modified cross-validation bandwidth. As for the variance of the modified cross-validation bandwidth, calculating the first-order term of $\mathrm{var}\left\{\widetilde{CV}'_n(\tilde{h}_{n0})\right\}$ will be enough, and so it will be useful to work with the simpler, linear approximation of $\hat{m}_h(x)$ given by (10).

## 3.1 Asymptotic results

The asymptotic bias and variance of the cross-validation bandwidth minimizing (11) are derived in this section. For this, some previous lemmas are proved. The detailed proof of these results can be found in `http://dm.udc.es/preprint/Barreiro_Ures_et_al_paper&SM.pdf`. The following assumptions are needed:

A1. $K$ is a symmetric and differentiable kernel function.

A2. For every $j = 0, \ldots, 6$, the integrals $\mu_j(K)$, $\mu_j(K')$ and $\mu_j(K^2)$ exist and are finite.

A3. The functions $m$ and $f$ are eight times differentiable.

A4. The function $\sigma^2$ is four times differentiable.

Lemma 3.1 provides expressions for the first and second order terms of both the bias and the variance of (9).

**Lemma 3.1** *Under assumptions* A1–A4, *the bias and the variance of the modified version of the Nadaraya–Watson estimator defined in* (9) *satisfy:*

$$\begin{aligned}\mathrm{E}\{\tilde{m}_h(x)\} - m(x) &= \mu_2(K)\left\{\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}h^2 \\ &+ \left[\mu_4(K)\left\{\frac{1}{24}m^{4)}(x) + \frac{1}{6}\frac{m'''(x)f'(x)}{f(x)} + \frac{1}{4}\frac{m''(x)f''(x)}{f(x)}\right.\right. \\ &+ \left.\left.\frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right\} - \mu_2(K)^2\frac{f''(x)}{f(x)}\left\{\frac{1}{4}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}\right]h^4 \\ &+ O\left(h^6 + n^{-1}\right)\end{aligned}$$

*and*

$$\begin{aligned}\mathrm{var}\{\tilde{m}_h(x)\} &= R(K)\sigma^2(x)f(x)^{-1}n^{-1}h^{-1} \\ &+ \left[\mu_2(K^2)f(x)^{-2}\left\{\varphi_3(x) + \frac{1}{2}m(x)^2f''(x) - 2\varphi_1(x)m(x)f(x)\right\}\right. \\ &- \left.R(K)\mu_2(K)\sigma^2(x)f(x)^{-2}f''(x)\right]n^{-1}h \\ &+ O(n^{-1}h^2 + n^{-2}h^{-2} + n^{-3}h^{-3}).\end{aligned}$$

It follows from Lemma 3.1 that

$$\tilde{M}_n(h) = B_1 h^4 + V_1 n^{-1} h^{-1} + B_2 h^6 + V_2 n^{-1} h + O\left(h^8 + n^{-1}h^2 + n^{-2}h^{-2} + n^{-3}h^{-3}\right),$$

where

$$\begin{aligned}
B_2 &= 2\mu_2(K) \int \left\{ \frac{1}{2} m''(x) + \frac{m'(x) f'(x)}{f(x)} \right\} \left[ \mu_4(K) \left\{ \frac{1}{24} m^{4)}(x) + \frac{1}{6} \frac{m'''(x) f'(x)}{f(x)} \right.\right. \\
&+ \left.\left. \frac{1}{4} \frac{m''(x) f''(x)}{f(x)} + \frac{1}{6} \frac{m'(x) f'''(x)}{f(x)} \right\} - \mu_2(K)^2 \frac{f''(x)}{f(x)} \left\{ \frac{1}{4} m''(x) + \frac{m'(x) f'(x)}{f(x)} \right\} \right] f(x)\, dx, \\
V_2 &= \int \left[ \mu_2(K^2) f(x)^{-2} \left\{ \frac{1}{2} f''(x) \sigma^2(x) + m'(x)^2 f(x) + \frac{1}{2} {\sigma^2}''(x) f(x) + f'(x) {\sigma^2}'(x) \right\} \right. \\
&- \left. R(K) \mu_2(K) \sigma^2(x) f(x)^{-2} f''(x) \right] f(x)\, dx.
\end{aligned}$$

are assumed to exist finite.

Lemma 3.2 provides expressions for the first and second order terms of both the expectation and variance of $\widetilde{CV}'_n(h)$.

**Lemma 3.2** *Let us define*

$$\begin{aligned}
A_1 &= 12\mu_2(K)\mu_4(K) \int f(x)^{-1} \left\{ \frac{1}{24} m^{(4)}(x) f(x) + \frac{1}{6} m'''(x) f'(x) + \frac{1}{4} m''(x) f''(x) \right. \\
&+ \left. \frac{1}{6} m'(x) f'''(x) \right\} \left\{ \frac{1}{2} m''(x) f(x) + m'(x) f'(x) \right\} dx \\
&- 6\mu_2(K)^3 \int f''(x) f(x)^{-2} \left\{ \frac{1}{2} m''(x) f(x) + m'(x) f'(x) \right\}^2, \\
A_2 &= \mu_2\left(K^2\right) \int f(x)^{-1} \left[ \frac{1}{2} f''(x) \sigma^2(x) + f'(x) (\sigma^2)'(x) \right. \\
&+ \left. f(x) \left\{ \frac{1}{2} (\sigma^2)''(x) + m'(x)^2 \right\} \right] dx \\
&- R(K)\mu_2(K) \int \sigma^2(x) f''(x) f(x)^{-1}\, dx, \\
R_1 &= 32 R(K)^2 \mu_2(K)^2 \int \sigma^2(x) f(x)^{-1} \left\{ \frac{1}{4} m''(x)^2 f(x)^2 + m'(x) m''(x) f(x) f'(x) \right. \\
&+ \left. m'(x)^2 f'(x)^2 \right\} dx, \\
R_2 &= 4\mu_2 \left\{ (K')^2 \right\} \int \sigma^2(x)^2\, dx.
\end{aligned}$$

*Then, under assumptions* A1–A4*, and assuming that $B_1$, $V_1$, $A_1$, $A_2$, $R_1$ and $R_2$ exist finite:*

$$\begin{aligned}
\mathrm{E}\left\{ \widetilde{CV}'_n(h) \right\} &= 4B_1 h^3 - V_1 n^{-1} h^{-2} + A_1 h^5 + A_2 n^{-1} + O\left(h^7 + n^{-1} h^2\right), \\
\mathrm{var}\left\{ \widetilde{CV}'_n(h) \right\} &= R_1 n^{-1} h^2 + R_2 n^{-2} h^{-3} + O\left(n^{-1} h^4 + n^{-2} h^{-1}\right).
\end{aligned}$$

Finally, Theorem 3.1, which can be derived from (13), Lemma 3.1 and Lemma 3.2, provides the asymptotic bias and variance of the cross-validation bandwidth that minimizes (11).

**Theorem 3.1** *Under the assumptions of Lemma 3.2 and assuming that $B_2$ and $V_2$ exist finite, the asymptotic bias and variance of the bandwidth that minimizes* (11) *are:*

$$\begin{aligned}
\mathrm{E}\left( \tilde{h}_{CV,n} \right) - \tilde{h}_{n0} &= \mathcal{B} n^{-3/5} + o\left(n^{-3/5}\right), \\
\mathrm{var}\left( \tilde{h}_{CV,n} \right) &= V n^{-3/5} + o\left(n^{-3/5}\right),
\end{aligned}$$

*where*

$$\begin{aligned} \mathcal{B} &= \frac{6B_2C_0^5 + V_2 - A_1C_0^5 - A_2}{12B_1C_0^2 + 2V_1C_0^{-3}}, \\ V &= \frac{R_1C_0^2 + R_2C_0^{-3}}{\left(12B_1C_0^2 + 2V_1C_0^{-3}\right)^2}. \end{aligned}$$

**Corollary 3.1** *Under the assumptions of Theorem 3.1, the asymptotic distribution of the bandwidth that minimizes* (11) *is given by:*

$$n^{3/10}\left(\tilde{h}_{CV,n} - \tilde{h}_{n0}\right) \xrightarrow{d} \mathrm{N}(0, V),$$

*where the constant $V$ was defined in Theorem 3.1.*

**Remark 3.1** *Although the results presented so far involve only the modified cross-validation bandwidth, defined as the bandwidth that minimizes* (11)*, it seems reasonable to think that these asymptotic results also apply to the standard cross-validation bandwidth defined in* (7)*, this being the rationale behind the decomposition of the Nadaraya–Watson estimator proposed in* (8)*. Under suitable assumptions, it can be proved that, as the sample size increases,*

$$\tilde{h}_{CV,n} - \tilde{h}_{n0} = \hat{h}_{CV,n} - h_{n0} + O_p\left(n^{-2/5}\right). \tag{14}$$

*Moreover, since $\tilde{h}_{n0} - h_{n0} = O\left(n^{-4/5}\right)$, it follows that*

$$\tilde{h}_{CV,n} = \hat{h}_{CV,n} + O_p\left(n^{-2/5}\right).$$

*A sketch of the proof of* (14) *and some other related results are included in* ***http://dm.udc.es/preprint/Barreiro_Ures_et_al_paper&SM.pdf***.

# 4 Bagged cross-validation bandwidth

While the cross-validation method is very useful to select reliable bandwidths in nonparametric regression, it also has the handicap of requiring a high computing time if the sample size is very large. This problem can be partially circumvented by using bagging (Breiman, 1996), a statistical technique belonging to the family of ensemble methods (Opitz and Maclin, 1999), in the bandwidth selection procedure. In this section, we explain how bagging may be applied in the cross-validation context. Additionally, the asymptotic properties of the corresponding selector are obtained. Apart from the obvious reductions in computing time, the bagging cross-validation selector also presents better theoretical properties than the leave-one-out cross-validation bandwidth. This will be corroborated in the numerical studies presented in Sections 6 and 7.

Let $\mathcal{X}^* = \{(X_1^*, Y_1^*), \ldots, (X_r^*, Y_r^*)\}$ be a random sample of size $r < n$ drawn without replacement from the i.i.d sample $\mathcal{X}$ defined in Section 2. This subsample is used to calculate a cross-validation bandwidth, $\hat{h}_{CV,r}$. A rescaled version of $\hat{h}_{CV,r}$, given by $(r/n)^{1/5}\hat{h}_{CV,r}$, can be viewed as a feasible estimator of the optimal MISE bandwidth, $h_{n0}$, for $\hat{m}_h$. Bagging consists of repeating this resampling procedure independently $N$ times, leading to $N$ rescaled bandwidths, $(r/n)^{1/5}\hat{h}_{CV,r,1}, \ldots, (r/n)^{1/5}\hat{h}_{CV,r,N}$. The bagging bandwidth is then defined as:

$$\hat{h}(r, N) = \frac{1}{N}\left(\frac{r}{n}\right)^{1/5} \sum_{i=1}^{N} \hat{h}_{CV,r,i}. \tag{15}$$

In the case of kernel density estimation, both the asymptotic properties and the empirical behavior of this type of bandwidth selector have been studied in Hall and Robinson (2009) for $N = \infty$ and generalized in Barreiro-Ures et al. (2020), where the asymptotic properties of the bandwidth selector are derived for the more practical case of a finite $N$. Furthermore, as discussed there, an alternative approach is to apply bagging to the cross-validation curves, wherein one

averages the cross-validation curves from $N$ independent resamples of size $r$, finds the minimizer of the average curve, and then rescales the minimizer as before. The asymptotic properties of the two approaches are equivalent, but we prefer bagging the bandwidths since doing so does not require as much communication between resamples and allows for parallel computing.

Following the same ideas employed in the previous section, a modified version of (15) can be defined. This modified bagged bandwidth uses modified cross-validation bandwidths $\tilde{h}_{CV,r,i}$ instead of $\hat{h}_{CV,r,i}$, for $i = 1, \ldots, N$, and it is given by

$$\tilde{h}(r, N) = \frac{1}{N} \left(\frac{r}{n}\right)^{1/5} \sum_{i=1}^{N} \tilde{h}_{CV,r,i}. \tag{16}$$

In the next section, the asymptotic bias and variance of the bagging bandwidth (16) when using the Nadaraya–Watson estimator (2) and the regression model (1) are obtained. Moreover, its asymptotic distribution is also derived. From these results and considering Remark 3.1, similar results for (15) could be obtained.

### 4.1 Asymptotic results

Expressions for the asymptotic bias and the variance of (16) are given in Theorem 4.1. The following additional assumption is needed:

A5. As $r, n \to \infty$, $r = o(n)$ and $N$ tends to a positive constant or $\infty$.

**Theorem 4.1** *Under assumptions* A1–A5*, the asymptotic bias and the variance of the bagged cross-validation bandwidth* $\tilde{h}(r, N)$ *are:*

$$\begin{aligned} \mathrm{E}\left\{\tilde{h}(r, N)\right\} - \tilde{h}_{n0} &= (\mathcal{B} + C_1) r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right), \\ \mathrm{var}\left\{\tilde{h}(r, N)\right\} &= V r^{-1/5} n^{-2/5} \left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\} + o\left(\frac{r^{-1/5} n^{-2/5}}{N} + r^{9/5} n^{-12/5}\right), \end{aligned}$$

*where the constants* $\mathcal{B}$ *and* $V$ *were defined in Theorem 3.1 and the constant* $C_1$ *is defined in expression* (48) *in* `http://dm.udc.es/preprint/Barreiro_Ures_et_al_paper&SM.pdf`.

**Corollary 4.1** *Under the assumptions of Theorem 4.1, the asymptotic distribution of the bagged cross-validation bandwidth* $\tilde{h}(r, N)$ *is:*

$$\frac{r^{1/10} n^{1/5}}{\sqrt{\frac{1}{N} + \left(\frac{r}{n}\right)^2}} \left\{\tilde{h}(r, N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0, V),$$

*where the constant* $V$ *was defined in Theorem 3.1. In particular, assuming that* $r = o\left(n/\sqrt{N}\right)$*, then,*

$$r^{1/10} n^{1/5} \sqrt{N} \left\{\tilde{h}(r, N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0, V).$$

Using (14) in Remark 3.1, it could be proved that similar results to those in Corollary 4.1 hold when considering $\hat{h}(r, N) - h_{n0}$ instead of $\tilde{h}(r, N) - \tilde{h}_{n0}$. It should be noted that, while $\hat{h}_{CV,n} - h_{n0}$ converges in distribution at the rate $n^{-3/10}$, this result can be improved with the use of bagging and letting $r$ and $N$ tend to infinity at adequate rates. For example, if both $r$ and $N$ tended to infinity at the rate $\sqrt{n}$, then $\hat{h}(r, N) - h_{n0}$ would converge in distribution at the rate $n^{-1/2}$, which is indeed a faster rate of convergence than $n^{-3/10}$.

## 5 Choosing an optimal subsample size

In practice, an important step of our approach is, for fixed values of $n$ and $N$, choosing the *optimal* subsample size, $r_0$. A possible optimality criterion, considering the modified bandwidths, could be

to select the value of $r$ that minimizes the main term of the variance of $\tilde{h}(r, N)$. In this case, we would get:

$$r_0^{(1)} = \frac{n}{3\sqrt{N}}$$

and the variance of the bagging bandwidth would converge to zero at the rate

$$\operatorname{var}\left\{\tilde{h}\left(r_0^{(1)}, N\right)\right\} \sim n^{-3/5} N^{-9/10},$$

which is a faster rate of convergence than that of the standard (modified) cross-validation bandwidth. In particular,

$$\frac{\operatorname{var}\left\{\tilde{h}\left(r_0^{(1)}, N\right)\right\}}{\operatorname{var}\left(\tilde{h}_{CV,n}\right)} \sim N^{-9/10}.$$

The obvious drawback of this criterion is that it would not allow any improvement in terms of computational efficiency, since the complexity of the algorithm would be the same as in the case of standard cross-validation, $O(n^2)$. This makes this choice of $r_0$ inappropriate for very large sample sizes. Another possible criterion for selecting $r_0$ would be to minimize, as a function of $r$, the asymptotic mean squared error (AMSE) of $\tilde{h}(r, N)$, given by:

$$\operatorname{AMSE}\left\{\tilde{h}(r, N)\right\} = (\mathcal{B} + C_1)^2 r^{-4/5} n^{-2/5} + V r^{-1/5} n^{-2/5} \left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\}. \tag{17}$$

Since $\mathcal{B}$, $C_1$ and $V$ are unknown, we propose the following method to estimate

$$r_0 = \arg\min_{r>1} \operatorname{AMSE}\left\{\tilde{h}(r, N)\right\}.$$

*Step 1.* Consider $s$ subsamples of size $p < n$, drawn without replacement from the original sample of size $n$.

*Step 2.* For each of these subsamples, obtain an estimate, $\hat{f}$, of the marginal density function of the explanatory variable (using kernel density estimation, for example) and an estimate, $\hat{m}$, of the regression function (for instance, by fitting a polynomial of a certain degree). Do the same for the required derivatives of both $f$ and $m$.

*Step 3.* Use the estimates obtained in the previous step to compute the constants $\mathcal{B}^{[i]}$, $C_1^{[i]}$ and $V^{[i]}$ for each subsample, where $i$ ($i = 1, \ldots, s$) denotes the subsample index.

*Step 4.* Compute the bagged estimates of the unknown constants, that is,

$$\hat{\mathcal{B}} = \frac{1}{s}\sum_{i=1}^{s} \mathcal{B}^{[i]}, \quad \hat{C}_1 = \frac{1}{s}\sum_{i=1}^{s} C_1^{[i]}, \quad \hat{V} = \frac{1}{s}\sum_{i=1}^{s} V^{[i]},$$

and obtain $\widehat{\operatorname{AMSE}}\left\{\tilde{h}(r, N)\right\}$ by plugging these bagged estimates into (17).

*Step 5.* Finally, estimate $r_0$ by:

$$\hat{r}_0 = \arg\min_{r>1} \widehat{\operatorname{AMSE}}\left\{\tilde{h}(r, N)\right\}.$$

Additionally, assuming that $r = o\left(n/\sqrt{N}\right)$, then

$$r_0^{(2)} = \left\{-\frac{4(\mathcal{B} + C_1)^2}{V} N\right\}^{5/3}$$

and the rate of convergence to zero of the AMSE of the bagging bandwidth would be:

$$\text{AMSE}\left\{\tilde{h}\left(r_0^{(2)}, N\right)\right\} \sim n^{-2/5} N^{-4/3}.$$

Hence,

$$\frac{\text{AMSE}\left\{\tilde{h}\left(r_0^{(2)}, N\right)\right\}}{\text{AMSE}\left(\tilde{h}_{CV,n}\right)} \sim n^{1/5} N^{-4/3},$$

and this ratio would tend to zero if $N$ tended to infinity at a rate faster than $n^{3/20}$. Furthermore, if we let $N = n^{3/20}$ and $r = r_0^{(2)}$, then the computational complexity of the algorithm would be $O\left(n^{13/20}\right)$, much lower than that of standard cross-validation. In fact, by selecting $r_0$ in this way, the complexity of the algorithm will only equal to that of standard cross-validation when $N$ tends to infinity at the rate $n^{6/13}$.

# 6 Simulation studies

The behavior of the leave-one-out and bagged cross-validation bandwidths is evaluated by simulation in this section. We considered the following regression models:

M1: $Y = m(X) + \varepsilon$, $m(x) = 2x$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

M2: $Y = m(X) + \varepsilon$, $m(x) = \sin(2\pi x)^2$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

M3: $Y = m(X) + \varepsilon$, $m(x) = x + x^2 \sin(8\pi x)^2$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

The Gaussian kernel was used for computing the Nadaraya–Watson estimator throughout this section. Moreover, to reduce computing time in the simulations, we used binning to select the ordinary and the bagged cross-validation bandwidths. The R (R Development Core Team, 2021) package `baggingbwsel` (Barreiro-Ures et al., 2021) was employed to carry out the simulation experiments.

Table 1: Ratio of the mean squared errors of the bagged and the ordinary cross-validation bandwidths for models M1–M3. Different values of $r$ and $N = 25$ were considered, for a sample size of $n = 10^5$.

| | Model | | |
|---|---|---|---|
| | M1 | M2 | M3 |
| Subsample size ($r$) | MSE ratio | | |
| 100 | 0.47 | 1.47 | 2.16 |
| 500 | 0.32 | 1.06 | 0.33 |
| 1,000 | 0.26 | 0.80 | 0.23 |
| 5,000 | 0.19 | 0.30 | 0.17 |
| 10,000 | 0.16 | 0.22 | 0.16 |

The effect that $r$ has on the mean squared error of the bagged bandwidth is also illustrated in Table 1, which shows the ratio of the mean squared errors of the bagged bandwidth and the ordinary cross-validation bandwidth, $\text{MSE}\{\hat{h}(r, N)\}/\text{MSE}(\hat{h}_{CV,n})$, for the three models.

Apart from a better statistical precision of the cross-validation bandwidths selected using bagging, another potential advantage of employing this approach is the reduction of computing times, especially with large sample sizes.

Table 2 shows the predicted CPU elapsed time for ordinary and bagged cross-validation for large sample sizes. Although we should take these predictions with caution, the results in Table 2 serve to illustrate the important reductions in computing time that bagging can provide for certain choices of $r$ and $N$, especially for very large sample sizes.

Much more detailed simulations can be found in `http://dm.udc.es/preprint/Barreiro_Ures_et_al_paper&SM.pdf`.

Table 2: Predicted CPU elapsed time for the standard and the bagging cross-validation method using three different choices for the subsample size.

| | Sample size ($n$) | | |
|---|---|---|---|
| | $10^6$ | $10^7$ | $10^8$ |
| Method | Computing time | | |
| Standard CV | 6 hours | 24 days | 7 years |
| Bagged CV ($r = n^{0.7}, N = 25$) | 40 seconds | 25 minutes | 16 hours |
| Bagged CV ($r = n^{0.8}, N = 25$) | 16 minutes | 17 hours | 45 days |
| Bagged CV ($r = n^{0.9}, N = 25$) | 3 hours | 11 days | 2 years |

# 7 Application to COVID-19 data

In order to illustrate the performance of the techniques studied in the previous sections, the COVID-19 dataset briefly mentioned in the introduction has been considered. It consists of a sample of size $n = 105,235$ which contains the age (the explanatory variable) and the time in hospital (the response variable) of people infected with COVID-19 in Spain from January 1, 2020 to December 20, 2020. A detailed analysis of this dataset can be found in `http://dm.udc.es/preprint/Barreiro_Ures_et_al_paper&SM.pdf`. For the sake of brevity, it is omitted here.

# Acknowledgments

This research has been supported by MINECO Grant MTM2017-82724-R, MICINN Grant PID2020-113578RB-I00, and by Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. The authors would like to thank the Spanish Center for Coordinating Sanitary Alerts and Emergencies for kindly providing the COVID-19 hospitalization dataset.

# References

Barbeito, I. (2020) *Exact bootstrap methods for nonparametric curve estimation.* Ph.D. thesis, Universidade da Coruña. `https://ruc.udc.es/dspace/handle/2183/26466`.

Barbeito, I., Cao, R. and Sperlich, S. (2021) Bandwidth selection for statistical matching and prediction. *Tech. rep.*, University of A Coruña. Department of Mathematics. `http://dm.udc.es/preprint/Bandwidth_Selection_Matching_Prediction_NOT_BLINDED.pdf` and `http://dm.udc.es/preprint/SuppMaterial_Bandwidth_Selection_Matching_Prediction_NOT_BLINDED.pdf`.

Barreiro-Ures, D., Cao, R., Francisco-Fernández, M. and Hart, J. D. (2020) Bagging cross-validated bandwidths with application to big data. *Biometrika*. `https://doi.org/10.1093/biomet/asaa092`.

Barreiro-Ures, D., Hart, J. D., Cao, R. and Francisco-Fernández, M. (2021) *baggingbwsel: Bagging Bandwidth Selection in Kernel Density and Regression Estimation.* R package version 1.0. `https://cran.r-project.org/package=baggingbwsel`.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Cao, R. and González-Manteiga, W. (1993) Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics*, **2**, 379–388.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Friedman, J. H. and Hall, P. (2007) On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, **137**, 669–683.

Hall, P. and Robinson, A. P. (2009) Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, **96**, 175–186.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83**, 86–95.

Köhler, M., Schindler, A. and Sperlich, S. (2014) A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review / Revue Internationale de Statistique*, **82**, 243–274.

Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability & Its Applications*, **9**, 141–142.

Opitz, D. and Maclin, R. (1999) Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, **11**, 169–198.

R Development Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Stone, C. J. (1977) Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–620. URL: `https://doi.org/10.1214/aos/1176343886`.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing*. London: Chapman and Hall.

Watson, G. S. (1964) Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 359–372.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# Un novo test de unimodalidade para datos circulares

Diego Bolón[1], Rosa M. Crujeiras[1] e Alberto Rodríguez-Casal[1]

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

**RESUMO**

As modas dunha poboación son puntos de alta frecuencia ao redor dos cales se acumula a maior parte da probabilidade. Na literatura estatística existen varios procedementos non paramétricos para contrastar o número de modas en datos na recta real. Pero a metodoloxía empregada na maioría destes tests impide que sexan directamente extensibles a outros contextos, como poden ser datos multidimensionais ou datos circulares. Partindo desta base, o principal obxectivo deste traballo é a construción dun novo test de multimodalidade para distribucións circulares, procurando ademais que sexa facilmente extensible a outros contextos, como poden ser o toro ou a esfera. Para iso, introducimos a idea de pseudo-verosimilitude como un análogo da función de verosimilitude da estatística paramétrica. Isto permítenos formular o noso test coa mesma estrutura dos tests de razón de verosimilitudes paramétricos apoiándonos no concepto de xanela crítica. Ademais, a pseudo-verosimilitude pode ser facilmente adaptada para facer inferencia en espazos de carácter xeral, pois para a súa construción só precisamos dun estimador non paramétrico da función de densidade da poboación. Unha vez proposto o novo test para contrastar o número de modas dunha poboación circular, comprobamos cal é o seu calibrado e potencia na práctica mediante un estudo de simulación. Para finalizar, ilustramos o funcionamento do test aplicándoo a un conxunto de datos reais.

**Palabras e frases chave**: Contraste; datos circulares; estimación tipo núcleo da densidade; multimodalidade; verosimilitude; xanela crítica.

## 1. INTRODUCIÓN E MOTIVACIÓN

Unha moda dun ángulo aleatorio absolutamente continuo $X$ é un punto $x_0 \in (0, 2\pi]$ onde a función de densidade de $X$, que denotaremos por $f$, ten un máximo local. Por tanto o concepto de moda fai referencia á idea de *concentración*: as modas son puntos de alta frecuencia ao redor dos cales se acumula a maior parte da probabilidade. Unha distribución (ou función de densidade) cunha soa moda denomínase *unimodal*. No caso de que teña máis dunha moda dise *multimodal*. Para tratar de realizar inferencia sobre o número de modas dunha poboación nacen os *tests de multimodalidade*. Dado $X$ un ángulo aleatorio absolutamente continuo con $j$ modas, un test de multimodalidade é un test estatístico que contrasta as hipóteses

$$H_0 \colon j \leq k \text{ fronte a } H_1 \colon j > k; \tag{1}$$

onde $k$ é un número natural fixado de antemán. Neste traballo centrarémonos no test que contrasta unimodalidade fronte a multimodalidade, é dicir

$$H_0 \colon j = 1 \text{ fronte a } H_1 \colon j > 1.$$

Coa intención de motivar a necesidade deste tipo de contrastes introducimos os datos analizados por Ameijeiras-Alonso et al. (2018). Esta base de datos contén todos os incendios detectados polo sensor de imaxe MODIS (acrónimo de *MODerate resolution Imaging Spectroradiometer*) da NASA

(*National Aeronautics and Space Administration*) en Galicia, dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012. En total rexistráronse 6804 incendios durante ese período, e a cada un deles asignóuselle un número do 1 ao 366 en función do día do ano no que comezou. Aquí, coñecer o número de modas, que neste contexto se corresponden coas tempadas de incencios ao longo do ano, cobra especial relevancia á hora de entender os seus patróns estacionais dos incendios forestais e así loitarmos mellor contra esta problemática.

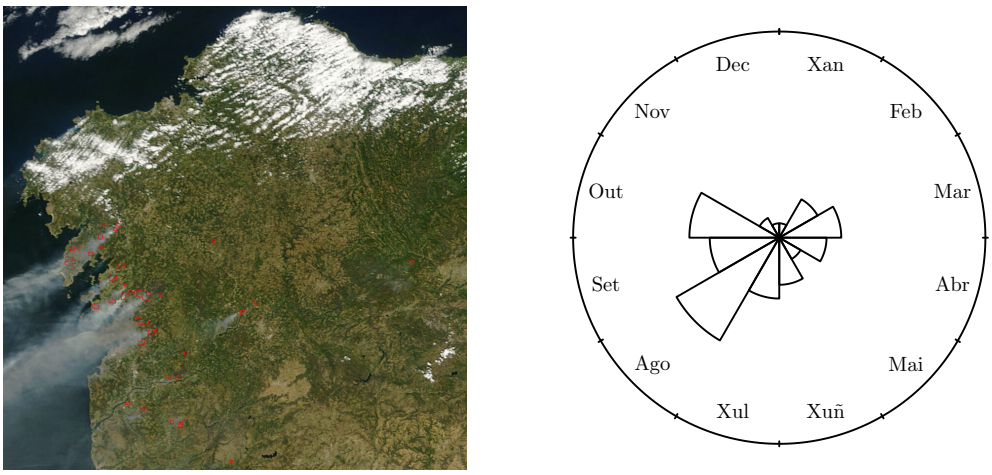



Figura 1: Á esquerda, os incendios (marcados en vermello) detectados polo satélite MODIS en Galicia o día 7 de agosto de 2006. Á dereita, o histograma circular (*rose diagram*) cos incendios detectados en Galicia dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012, detectados polo satélite MODIS da NASA.

Na Figura 1 pódense ver os datos anteriores representados nun histograma circular. Neste gráfico dividimos a circunferencia en doce rexións, cada unha correspondente a un mes do ano. Despois engadimos un sector circular a cada unha das seccións, de xeito que a área de cada sector sexa directamente proporcional a cantidade de incendios detectada nese mes. No histograma circular vemos que, como era de esperar, a maior parte dos incendios detectados sucederon ao longo do verán, entre os meses de xullo e setembro. Pero tamén parece haber outros dous períodos con alta frecuencia de incendios; un a principios do outono, principalmente en outubro, e outro a finais do inverno, en marzo. Estes dous picos de incendios están situados fóra da parte do ano con condicións meteorolóxicas máis favorables para a formación natural de incendios, polo que poden estar asociados a certos comportamentos humanos, como a queima preventiva ou a queima de restrollos. Así, se tivesemos un test de multimodalidade que nos permita afirmar que hai probas estatísticamente significativas para a existencia de máis dunha moda con esta mostra estaríamos aportando evidencia de que a actividade humana condiciona a estrutura das vagas de incendios ao longo do ano.

Tendo en conta o anterior, o noso obxectivo é construír un test de multimodalidade para datos circulares que conte cunha metodoloxía facilmente adaptable a outros contextos, como poden ser datos direccionais, multidimensionais, no toro...

A estrutura deste traballo é a seguinte. Na Sección 2 comezamos introducindo brevemente as principais características do test de razón de verosimilitudes paramétrico, e apoiándonos nelas construímos a nosa proposta de test de multimodalidade para datos circulares. Na Sección 3 falamos das propiedades do selector de xanela $h_{max}$, que teñen consecuencias no comportamento do noso test. Na Sección 4 realizamos dous estudos de simulación para comprobar o calibrado e potencia do test na práctica e comentamos os resultados obtidos. Na Sección 5 aplicamos o test aos datos presentados nesta introdución e finalmente na Sección 6 comentamos as principais conclusións do traballo e como se podería adaptar esta nova proposta de test de multimodalidade a outros espazos que non sexan a circunferencia. A proba da Proposición 1, que precisamos para a construción do test, está recollida na Sección 7.

## 2. PROPOSTA DE TEST

Para abordar o problema (1), propoñemos un test de multimodalidade baseado no test de razón de verosimilitudes da estatística paramétrica. O test de razón de verosimilitudes, introducido por primeira vez por Neyman e Pearson no ano 1936, é unha familia de contrastes estatísticos que se empregan en inferencia paramétrica para contrastar a situación do parámetro descoñecido dentro espazo de parámetros. Supoñamos que $X_1, X_2, \ldots, X_n$ é unha mostra aleatoria simple dunha variable aleatoria absolutamente continua con función de densidade $f$ pertencente a familia paramétrica $\{f_\theta \colon \theta \in \Theta\}$, onde $\Theta \subset \mathbb{R}^m$. Dividamos o espazo de parámetros en dous subconxuntos disxuntos, escollendo $\Theta_0$ e $\Theta_1$ tales que $\Theta_0 \cap \Theta_1 = \emptyset$ e $\Theta_0 \cup \Theta_1 = \Theta$. O test de razón de verosimilitudes contrasta hipóteses da forma

$$H_0 \colon \theta \in \Theta_0 \text{ fronte a } H_1 \colon \theta \in \Theta_1.$$

Para a construción do estatístico de contraste, este tipo de tests emprega a función de verosimilitude $\mathcal{L}$, que é a función real positiva definida como

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

A idea detrás da verosimilitude é que $\mathcal{L}(\theta_0)$ representa a *factibilidade* de que o valor do parámetro descoñecido sexa $\theta_0$ unha vez observada a mostra. Así, dada unha mostra, que $\mathcal{L}(\theta_0)$ sexa maior que $\mathcal{L}(\theta_1)$ significa que é máis *verosímil* que o parámetro descoñecido $\theta$ sexa igual a $\theta_0$ que a $\theta_1$ á vista dos datos observados. Entón, no caso de que a hipótese nula sexa certa, a función de verosimilitude debería tomar valores grandes dentro do conxunto $\Theta_0$, que é o que se corresponde a $H_0$. Polo tanto, un candidato a estatístico de contraste sería

$$\lambda = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \in [0, 1];$$

que estará ben definido se a verosimilitude $\mathcal{L}$ está limitada en todo o espazo de parámetros $\Theta$. Tendo en conta o razoamento anterior, o test de razón de verosimilitudes rexeitará a hipótese nula para valores pequenos do estatístico $\lambda$.

En vez de empregar $\lambda$, adóitase utilizar o estatístico equivalente

$$D = -2\log(\lambda) = 2\left[\sup_{\theta \in \Theta} \ell(\theta) - \sup_{\theta \in \Theta_0} \ell(\theta)\right] \geq 0,$$

onde $\ell(\theta) = \log \mathcal{L}(\theta)$, rexeitando agora a hipótese nula para valores grandes de $D$. Isto permítenos obter un test asintótico, pois o Teorema de Wilks (Wilks, 1938) asegura que, baixo a hipótese nula, $D$ converxe en distribución a unha chi cadrado baixo condicións de regularidade bastante laxas. Ademais, os contrastes de razón de verosimilitudes resultan ser os tests uniformemente máis potentes en varios escenarios, tal e como garanten resultados como o Lema de Neyman-Person (Neyman e Pearson, 1933), o Teorema de Karlin-Rubin ou o Teorema de Lehmann (Karlin, 1957). Todo o anterior explica a gran popularidade do test de razón de verosimilitudes dentro da estatística paramétrica. Baseando o novo test de multimodalidade no test de razón de verosimilitudes buscamos que herde estas boas características, así como conseguir unha metodoloxía adaptable para traballar en calquera espazo.

Para construír un análogo non paramétrico da función de verosimilitude apoiámonos no estimador tipo núcleo da densidade. Sexa $X_1, \ldots, X_n$ unha mostra aleatoria simple do ángulo aleatorio $X$. O estimador tipo núcleo da función de densidade de $X$ é a función

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(x - X_i\right); \tag{2}$$

onde $h \in \Theta = (0, +\infty)$ é un número real positivo, e $K_h$ é unha familia de densidades circulares simétricas e centradas no ángulo cero indexadas polo parámetro $h$. As funcións $K_h$ denomínanse

*funcións núcleo* ou *kernels*, mentres que $h$ recibe o nome de *xanela* ou *parámetro de suavizado* (*bandwidth* en inglés). Unha escolla habitual da función núcleo é a normal enrolada $\mathrm{WN}(0, h^2)$:

$$K_h(x) = \frac{1}{\sqrt{2\pi h^2}} \sum_{k=-\infty}^{+\infty} \exp\left(\frac{-(x+2\pi k)^2}{2h^2}\right). \tag{3}$$

Unha vez obtida unha estimación da densidade de $X$ podemos derivar desta un análogo da función de verosimilitude paramétrica. Así, definimos a *pseudo-verosimilitude* da mostra como a función real positiva

$$\mathcal{L}(h) = \prod_{i=1}^{n} \hat{f}_h(X_i), \tag{4}$$

onde $h > 0$. Poderíamos pensar na función $\mathcal{L}$ como nunha verosimilitude paramétrica onde a familia paramétrica de densidades que estamos a supoñer é $\{\hat{f}_h \colon h > 0\}$, aínda que esta familia non é independente dos datos observados, senón que vén determinada directemente por eles.

Para poder formular o test de multimodalidade coa linguaxe dos tests de razón de verosimilitudes paramétricos temos traducir as hipóteses a contrastar definidas en (1) nunha división do espazo de parámetros $\Theta = (0, +\infty)$. Isto lograrémolo apoiándonos de novo na estimación de núcleo da densidade. Así, os valores de $h$ que ofrezan unha estimación tipo núcleo con $k$ modas como máximo serán os asociados a hipótese nula e os demais estarán asociados a hipótese alternativa. Esta división do espazo de parámetros pódese simplificar tendo en conta que o número de modas do estimador tipo núcleo $\hat{f}_h$ é unha función decrecente en $h$ sempre que empreguemos como núcleo a normal enrolada (véxase Huckemann et al., 2016). Isto permítenos definir o concepto de xanela crítica. A xanela crítica para $k$ modas, $h_k$, non é máis que o menor valor do parámetro de suavizado para o cal a estimación tipo núcleo da densidade ten $k$ modas como máximo. É dicir:

$$h_k = \text{mín}\{h > 0 \colon \hat{f}_h \text{ ten } k \text{ modas}\}.$$

Así, $h_k$ representa a fronteira entre as estimacións tipo núcleo da densidade con máis e menos de $k$ modas: se o parámetro de suavizado $h$ é menor que $h_k$, entón $\hat{f}_h$ ten máis de $k$ modas, e se $h$ é maior que $h_k$, entón o número de modas de $\hat{f}_h$ non é maior que $k$. Pero esta é precisamente a división entre a hipótese nula e a hipótese alternativa que definimos en (1). Por tanto, a xanela crítica facilítanos enormemente a división do espazo de parámetros $\Theta$:

$$\Theta_0 = \{h > 0 \colon \hat{f}_h \text{ ten } k \text{ modas como máximo}\} = [h_k, +\infty);$$
$$\Theta_1 = \{h > 0 \colon \hat{f}_h \text{ ten máis de } k \text{ modas}\} = (0, h_k).$$

Xa temos introducidas todas as ferramentas para poder construír o noso test de multimodalidade. O estatístico de contraste, será:

$$D_k = 2\left[\sup_{h>0} \ell(h) - \sup_{h \geq h_k} \ell(h)\right],$$

onde $\ell(h) = \log \mathcal{L}(h)$. Igual que no test de razón de verosimilitudes usual, rexeitamos a hipótese nula para valores grandes de $D_k$.

Pero, tal como está pensado, o estatístico de contraste $D_k$ non está ben definido. Resulta sinxelo ver que $\lim_{h\to 0} \mathcal{L}(h) = +\infty$, polo que $\mathcal{L}$ non é unha función limitada e por tanto $\sup_{h>0} \ell(h)$ non é un número real. Un xeito de solventar este problema é redefinir a función $\mathcal{L}$ dada por (4) mediante validación cruzada. Para iso, imos apoiarnos nas funcións auxiliares:

$$\hat{f}_h^{-i}(x) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} K_h\left(x - X_j\right).$$

Daquela, definimos a *pseudo-verosimilitude por validación cruzada* da mostra $X_1, \ldots, X_n$ como

$$\mathcal{L}_{CV}(h) = \prod_{i=1}^{n} \hat{f}_h^{-i}(X_i), \quad h > 0. \tag{5}$$

Con esta modificación logramos que a función $\mathcal{L}_{CV}$ si que esté limitada superiormente, tal como nos garante o seguinte resultado.

**Proposición 1.** *Sexa $X_1, \ldots, X_n$ unha mostra aleatoria simple dun ángulo aleatorio absolutamente continuo $X$. Sexa $\mathcal{L}_{CV}$ a función de pseudo-verosimilitude por validación cruzada definida en* (5)*, onde $K_h$ é a normal enrolada definida en* (3)*. Entón a función $\mathcal{L}_{CV}(h)$ está limitada superiormente no intervalo $(0, +\infty)$ con probabilidade 1.*

A proba deste resultado recóllese na Sección 7. Polo tanto, o estatístico

$$D_k = 2\left[\max_{h>0} \ell_{CV}(h) - \max_{h\geq h_k} \ell_{CV}(h)\right],$$

onde $\ell_{CV}(h) = \log \mathcal{L}_{CV}(h)$, está ben definido pola Proposición 1. Este será finalmente o estatístico de contraste que empreguemos no noso test de multimodalidade. Igual que no test de razón de verosimilitudes usual, rexeitaremos a hipótese nula para valores grandes de $D_k$.

Para o calibrado empregamos bootstrap suavizado, xerando remostras da densidade suavizada $\hat{f}_{h_k}$, onde $k$ é precisamente o número de modas que queremos contrastar. A partir destas remostras calculamos réplicas do estatístico $D_k$ e estimaremos o p-valor do test calculando a proporción de réplicas maiores que o valor do estatístico sobre a mostra orixinal.

Así, de xeito esquemático, o test de multimodalidade con nivel de significación $\alpha \in (0, 1)$ para datos circulares é:

1. A partir da mostra $X_1, \ldots, X_n$, obtemos a xanela crítica

   $$h_k = \text{mín}\{h > 0\colon \hat{f}_h \text{ ten } k \text{ modas}\}$$

   e as xanelas que maximizan a verosimilitude por validación cruzada baixo a hipótese nula e a alternativa:

   $$\mathcal{L}_{CV}(h_{max}) = \max_{h>0}\{\mathcal{L}_{CV}(h)\}; \quad \mathcal{L}_{CV}(h_{H_0}) = \max_{h\geq h_k}\{\mathcal{L}_{CV}(h)\}.$$

   E con elas calculamos o estatístico de contraste:

   $$D_k = 2\left[\ell_{CV}(h_{max}) - \ell_{CV}(h_{H_0})\right].$$

2. Obtemos a remostra $X_1^*, \ldots, X_n^*$ da densidade suavizada $\hat{f}_{h_k}$ e calculamos o valor do estatístico para a remostra, que denotamos por $D_k^*$.

3. Repetimos $B$ veces o paso 2, conseguindo así B réplicas do estatístico: $D_k^{*,1}, D_k^{*,2}, \ldots, D_k^{*,B}$.

4. O test rexeitará a hipótese nula de que $f$ ten como máximo $k$ modas se

   $$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(D_k^{*,b} > D_k) < \alpha.$$

## 3. O SELECTOR DE XANELA $h_{max}$

Tamén podemos empregar a función de pseudo-verosimilitude para seleccionar un parámetro de suavizado $h$ á hora de construír o estimador da densidade $\hat{f}_h$ que definimos en (2). Dada a mostra circular $X_1, \ldots, X_n$, poderíase escoller como xanela o valor $h_{max} > 0$ que verifique que

$$\mathcal{L}_{CV}(h_{max}) = \max_{h>0} \mathcal{L}_{CV}(h);$$

que existe por ser a función $\mathcal{L}_{CV}$ limitada superiormente.

As principais propiedades deste selector da xanela foron estudadas por Hall et al. (1987), que probaron que $h_{max}$ é un valor *óptimo* do parámetro de suavizado, no sentido de que o estimador tipo núcleo $\hat{f}_{h_{max}}$ é un estimador asintóticamente consistente da verdadeira función de densidade $f$, sempre que $f$ sexa o suficientemente regular e esté limitada fóra de cero (para máis información sobre as condicións de regularidade necesarias, véxase Hall et al., 1987). $h_{max}$ é, polo tanto, unha escolla habitual do parámetro de suavizado á hora de estimar a función de densidade dun ángulo aleatorio mediante unha estimación tipo núcleo.

Ademais, as condicións necesarias por Hall et al. (1987) para a converxencia asintótica de $\hat{f}_{h_{max}}$ adiántannos unha das posibles fraquezas do novo contraste de multimodalidade para datos circulares. Como precisamos que $f$ esté limitada fóra de 0 para garantir a consistencia do estimador $\hat{f}_{h_{max}}$, é esperable que o test teña problemas ao detectar o verdadeiro número de modas dunha distribución cando a densidade desta se anula, ben nun punto, ben nun arco de circunferencia de medida positiva. Daquela, deberemos de prestarlle especial atención a este tipo de situacións á hora de comprobar o rendemento de test na práctica.

## 4. ESTUDO DE SIMULACIÓN E RESULTADOS

Nesta sección estudaremos o comportamento na práctica da nova proposta de test de multimodalidade mediante dous estudos de simulación. No primeiro deles buscamos comprobar o calibrado do novo test á hora de contrastar a unimodalidade dos datos, e no segundo estudaremos a potencia do mesmo ante unha alternativa de bimodalidade.

En ambos os estudos de simulación xeráronse $M = 1000$ mostras de tamaño $n$ de diversas distribucións. A cada unha destas mostras aplicámoslle o novo test, vendo se rexeita ou non a unimodalidade dos datos para varios niveis de significación $\alpha$. Finalmente, calulamos a proporción de mostras rexeitadas do total. Imos realizar o anterior para dous tamaños de mostra distintos, $n = 100$ e $n = 500$, tanto para estudar potencia como calibrado. Os niveis de significación considerados son os tres máis usuais: $\alpha = 0.01$, $\alpha = 0.05$ e $\alpha = 0.1$, e os p-valores aproxímanse mediante $B = 500$ remostra bootstrap en todos os casos. Para datos comprobar o calibrado empréganse cinco distribucións unimodais distintas, mentres para o estudo de potencia considéranse catro distribucións bimodais diferentes. As distribucións unimodais consideradas foron:

- **Modelo 1 (M1)**: unha distribución de von Mises: $\text{vM}(0, 1)$.
- **Modelo 2 (M2)**: unha mixtura de von Mises: $0.2 \cdot \text{vM}(2\pi/3, 3) + 0.6 \cdot \text{vM}(\pi, 1.4) + 0.2 \cdot \text{vM}(4\pi/3, 3)$.
- **Modelo 3 (M3)**: unha mixtura de von Mises: $0.05 \cdot \text{vM}(2\pi/3, 7) + 0.9 \cdot \text{vM}(\pi, 1) + 0.05 \cdot \text{vM}(4\pi/3, 7)$.
- **Modelo 4 (M4)**: unha von Mises *sine-skewed*: $\text{ssvM}(\pi, 1, -0.9)$.
- **Modelo 5 (M5)**: unha distribución beta modificada para que o soporte sexa o intervalo $[\pi/2, 3\pi/2]$ e enrolada: $\exp[i\,(\pi \cdot \text{Beta}(3, 2) + \pi/2)]$.

Estes modelos buscan representar a unha gran variedade de situacións, empezando por casos simples (unha von Mises, Modelo 1), pasando por modas planas (Modelo 2), distintos graos de asimetría (Modelos 4 e 5) e finalizado cunha distribución sectorial, onde todos os datos están concentrados na semicircunferencia esquerda (Modelo 5).

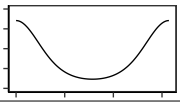
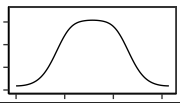
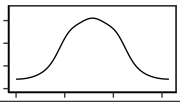
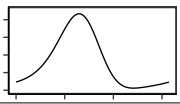
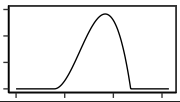
Por outro lado, as distribucións circulares con dúas modas escollidas para o estudo de potencia son:

- **Modelo 6 (M6)**: unha mixtura de dúas von Mises: $0.5 \cdot \text{vM}(2, 5) + 0.5 \cdot \text{vM}(4, 5)$.
- **Modelo 7 (M7)**: unha mixtura de von Mises: $0.9 \cdot \text{vM}(\pi/2, 2) + 0.1 \cdot \text{vM}(3\pi/2, 3)$.
- **Modelo 8 (M8)**: unha mixtura de von Mises: $0.5 \cdot \text{vM}(\pi - 1, 1.5) + 0.5 \cdot \text{vM}(\pi + 1, 1.5)$.
- **Modelo 9 (M9)**: unha mixtura de tres von Mises: $0.3 \cdot \text{vM}(\pi/2, 6) + 0.5 \cdot \text{vM}(3\pi/4, 2) + 0.2 \cdot \text{vM}(7\pi/4, 4)$.

Igual que cos modelos unimodais, estas distribucións buscan representar unha gran variedade de situacións: modas separadas (Modelo 6), modas xuntas (Modelo 8), unha moda secundaria moito menor que a principal (Modelos 7 e 9), e ata un modelo con certa asimetría (Modelo 9).

Para a realización do novo test de multimodalidade empregouse código propio deseñado *ex professo* para este estudo de simulación. Debido ao seu alto custo computacional, todas as simulacións foron realizadas mediante os recursos computacionais do Centro de Supercomputación de Galicia (CESGA).

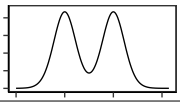
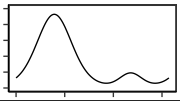
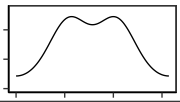
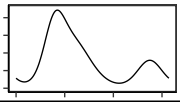
Táboa 1: Proporcións de rexeitamento para o test de multimodalidade baseado en pseudo-verosimilitude con nivel de significación ao 1 %, 5 % e 10 % calculadas a partir de $M = 1000$ mostras. Ao lado de cada proporción de rexeitamento, e entre paréntese, aparece a súa desviación típica estimada multiplicada por 1.96. O test calibrouse empregando $B = 500$ remostras. Cada fila correspóndese cunha combinación de distribución e tamaño de mostra distinta.

| Modelo | Densidade | Tamaño | Nivel de significación | | |
|---|---|---|---|---|---|
| | | | 1 % | 5 % | 10 % |
| M1 |  | $n = 100$ | 0.001(0.002) | 0.013(0.007) | 0.038(0.012) |
| | | $n = 500$ | 0.005(0.004) | 0.033(0.011) | 0.065(0.015) |
| M2 |  | $n = 100$ | 0.003(0.003) | 0.036(0.012) | 0.072(0.016) |
| | | $n = 500$ | 0.015(0.008) | 0.053(0.014) | 0.101(0.019) |
| M3 |  | $n = 100$ | 0.004(0.004) | 0.027(0.010) | 0.057(0.014) |
| | | $n = 500$ | 0.014(0.007) | 0.049(0.013) | 0.095(0.018) |
| M4 |  | $n = 100$ | 0.011(0.006) | 0.047(0.013) | 0.075(0.016) |
| | | $n = 500$ | 0.009(0.006) | 0.041(0.012) | 0.081(0.017) |
| M5 |  | $n = 100$ | 0.011(0.006) | 0.063(0.015) | 0.133(0.021) |
| | | $n = 500$ | 0.030(0.011) | 0.138(0.021) | 0.232(0.026) |

Comezamos comprobando o calibrado do test a hora de detectar unimodalidade. Na Táboa 1 aparecen os resultados deste estudo de simulación, e pódese observar que o novo test parece estar ben calibrado para tamaños de mostra grandes. O contraste só presenta proporcións de rexeitamento superiores ao nivel de significación para o Modelo 5. Corrobórase por tanto a sospeita que expuxemos na sección anterior e o test ten problemas á hora de detectar a hipótese nula cando a función de densidade dos datos se anula nun sector circular de lonxitude positiva. Para os Modelos 2, 3 e 4 o test é conservador para mostras pequenas ($n = 100$) na gran maioría dos casos, con proporcións de rexeitamento menores ao nivel de significación. As únicas excepcións están no Modelo 4 con $\alpha = 0.01$ e $\alpha = 0.05$, onde as proporcións de rexeitamento están moi próximas ao nivel nos dous casos. Pola contra, para mostras de tamaño $n = 500$ o test ofrece proporcións de rexeitamento similares ao nivel nestes tres modelos, salvo no caso do Modelo 4 con $\alpha = 0.1$, onde a proporción de rexeitamento está lixeiramente por debaixo do nivel. Por último, para o Modelo 1 o test é conservador para todos os tamaños e niveis de significación, é dicir, as proporcións de rexeitamento están por debaixo do nivel para todos os casos.

Pasamos logo a estudar a potencia do test á hora de detectar a alternativa de bimodalidade. Os resultados deste segundo estudo de simulación aparecen recollidos na Táboa 2. O test parece detectar satisfactoriamente a alternativa en todos os escenarios considerados. Ademais, os resultados son coherentes co esperado tendo en conta a forma das distribucións empregadas. O test logra proporcións de rexeitamento altas para os modelos con modas claramente separadas, e ademais as proporcións de rexeitamento ván diminuíndo ao diminuír o tamaño da moda secundaria (por orde, de maior a menor, Modelo 6, Modelo 9 e Modelo 7). O test logra proporcións de rexeitamento moito máis pequenas para a distribución con modas pegadas (Modelo 8), pero sempre maiores que o nivel de significación. Por outra parte, as proporcións de rexeitamento crecen ao aumentar o tamaño da mostra en todos os casos.

Táboa 2: Proporcións de rexeitamentos para o test baseado na pseudo-verosimilitude con nivel de significación ao 1 %, 5 % e 10 % calculadas a partir de $M = 1000$ mostras. Ao lado de cada proporción de rexeitamento, e entre paréntese, aparece a súa desviación típica estimada multiplicada por 1.96. O test calibrouse empregando $B = 500$ remostras. Cada fila correspóndese cunha combinación de distribución e tamaño de mostra distinta.

| Modelo | Densidade | Tamaño | Nivel de significación | | |
|---|---|---|---|---|---|
| | | | 1 % | 5 % | 10 % |
| M6 |  | $n = 100$ | 0.911(0.018) | 0.996(0.004) | 0.997(0.003) |
| | | $n = 500$ | 1.000(0.000) | 1.000(0.000) | 1.000(0.000) |
| M7 |  | $n = 100$ | 0.295(0.028) | 0.575(0.031) | 0.714(0.028) |
| | | $n = 500$ | 0.985(0.008) | 0.995(0.004) | 0.996(0.004) |
| M8 |  | $n = 100$ | 0.033(0.011) | 0.095(0.018) | 0.131(0.021) |
| | | $n = 500$ | 0.106(0.019) | 0.256(0.027) | 0.369(0.030) |
| M9 |  | $n = 100$ | 0.373(0.030) | 0.684(0.029) | 0.820(0.024) |
| | | $n = 500$ | 1.000(0.000) | 1.000(0.000) | 1.000(0.000) |

## 5. APLICACIÓN A DATOS REAIS

Unha vez comprobado o bo calibrado do novo test de multimodalidade para datos circulares, podemos aplicalo aos datos de incendios de Ameijeiras-Alonso et al. (2018) presentados na introdución para ver se podemos afirmar que existe máis dunha moda neste caso. Calculamos o estatístico de contraste para a mostra dos incendios é obtemos o valor $D_1 = 8142.012$. Estimamos o p-valor asociado a mostra mediante $B = 500$ remostras, e a aproximación conseguida do p-valor é 0. Polo tanto, o contraste rexeita a unimodalidade dos datos baixo calquera nivel de significación usual. Ou o que é o mesmo, hai probas estatisticamente significativas para afirmar que hai máis dunha tempada de incendios ao longo do ano, algo que, como indicamos na introdución, podería verse explicado pola influencia da actividade humana no patrón natural de incendios en Galicia.

## 6. CONCLUSIÓNS

A principal conclusión deste traballo é clara. O novo test de multimodalidade para datos circulares baseado na pseudo-verosimilitude está ben calibrado e detecta satisfactoriamente a hipótese altertiva. Ademais, para a construción do mesmo só precisamos do estimador tipo núcleo da función de densidade $\hat{f}_h$. Este tipo de estimadores son facilmente adaptables a distintos tipos de espazos, como poden ser a esfera (véxase Mardia e Jupp, 2000, Cáp. 12), o cilindro ou o toro. Por tanto, o noso test de multimodalidade será extensible de forma directa a todos estes espazos onde o estimador tipo núcleo está ben definido. Iso si, a necesidade de que as funcións de densidade estén limitadas fóra de cero apunta a que o test só vaia ter un bo comportamento en variedades compactas (esfera, toro...), pois en variedades non compactas esta restricción non se pode verificar. A principal problemática de adaptar o test a estes espazos está no calibrado do mesmo, pois en máis dunha dimensión non hai polo de agora un concepto paralelo á xanela crítica unidimensional. Por tanto, o noso calibrado mediante bootstrap suavizado, que se apoia plenamente na xanela crítica, debe de ser reformulado.

## 7. PROBA DA PROPOSICIÓN 1

*Demostración da Proposición 1.* Sexa $K_h$ a función de densidade dunha normal enrolada $\mathrm{WN}(0, h^2)$ definida en (3). Tendo en conta que $\lim_{h\to+\infty} K_h(x) = 1/2\pi$ para todo $x \in \mathbb{R}$, temos que:

$$\lim_{h\to+\infty} \mathcal{L}_{CV}(h) = \lim_{h\to+\infty} \prod_{i=1}^{n} \hat{f}_h^{-i}(X_i) = (2\pi)^{-n}.$$

Imos ver un resultado similar pero cando $h \to 0$. Para iso temos que traballar coa definición da normal enrolada $\mathrm{WN}(0, h^2)$:

$$K_h(x) = \frac{1}{\sqrt{2\pi h^2}} \sum_{k\in\mathbb{Z}} \exp\left(-\frac{(x-2\pi k)^2}{2h^2}\right) \tag{6}$$

Imos estudar a serie en (6) separando os termos positivos dos negativos. Daquela, para os positivos temos que:

$$\sum_{k=0}^{+\infty} \exp\left(-\frac{(x+2\pi k)^2}{2h^2}\right) \leq \sum_{k=0}^{+\infty} \exp\left(-\frac{x^2}{2h^2} - \frac{2\pi^2 k^2}{h^2}\right) \leq$$
$$\leq \exp\left(-\frac{x^2}{2h^2}\right) \sum_{k=0}^{+\infty} \exp\left(-\frac{2\pi^2 k}{h^2}\right) = \exp\left(-\frac{x^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}. \tag{7}$$

A primeira desigualdade de (7) dedúcese de que $(a+b)^2 \geq a^2+b^2$ se $a, b \geq 0$. A última desigualdade é inmediata, pois estámoslle a sumar máis términos á serie (hai máis números naturais que cadrados perfectos).

Aplicando un razonamento similar para os termos negativos, temos que:

$$\sum_{k=1}^{+\infty} \exp\left(-\frac{(x-2\pi k)^2}{2h^2}\right) \leq \exp\left(-\frac{(2\pi-x)^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}.$$

De todo o anterior deducimos que:

$$K_h(x) \leq \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1} \left[\frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(2\pi-x)^2}{2h^2}\right) + \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{x^2}{2h^2}\right)\right];$$

e por tanto $\lim_{h\to 0} K_h(x) = 0$ para todo $x \in (0, 2\pi)$.

Supoñamos agora que todos os valores $X_1, \ldots, X_n$ son todos distintos entre si. Entón, o anterior implica

$$\lim_{h\to 0} \hat{f}_h^{-i}(X_i) = \lim_{h\to 0} \frac{1}{(n-1)} \sum_{j=1, j\neq i}^{n} K_h\left(X_i - X_j\right) = 0,$$

para todo $i \in \{1, \ldots, n\}$ e por tanto $\lim_{h\to 0} \mathcal{L}_{CV}(h) = 0$. Como $\mathcal{L}_{CV}(h)$ é continua en $(0, +\infty)$, entón temos que $\mathcal{L}_{CV}(h)$ está limitada no intervalo $(0, +\infty)$ sempre que os valores $X_1, \ldots, X_n$ sexan todos distintos entre si. Como $X$ é un ángulo aleatorio absolutamente continuo, isto sucede con probabilidade 1. □

## AGRADECEMENTOS

Debo expresar a miña gratitude co Centro de Supercomputación de Galicia (CESGA), pois os seus recursos computacionais foron imprescindibles para a realización de todos os estudos de simulación deste traballo. Tamén debo dar as grazas a Jose Ameijeiras Alonso por facilitarnos os datos dos incendios presentados na introdución.

Este traballo foi realizado grazas ao apoio económico do Proxecto MTM2016–76969–P (INN-PAR2D) da Axencia Estatal de Investigación (AEI) cofinanciado polo Fondo Europeo de Desenvolvemento Rexional (ERDF).

## REFERENCIAS

Ameijeiras-Alonso, J., Crujeiras, R. M., e Rodríguez-Casal, A. (2018). Directional statistics for wildfires. En *Applied Directional Statistics*, páx. 203–226. Chapman and Hall/CRC.

Hall, P., Watson, G., e Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762.

Huckemann, S., Kim, K.-R., Munk, A., Rehfeldt, F., Sommerfeld, M., Weickert, J., Wollnik, C., et al. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli*, 22(4):2113–2142.

Karlin, S. (1957). Pólya type distributions, II. *The Annals of Mathematical Statistics*, 28(2):281–308.

Mardia, K. V. e Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons.

Neyman, J. e Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# MAXIMUM LIKELIHOOD ESTIMATION IN SINGLE-INDEX MIXTURE CURE MODELS

Beatriz Piñeiro-Lamas[1], Ana López-Cheda[1] and Ricardo Cao[1,2]

[1] Grupo MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña.
[2] Instituto Tecnolóxico de Matemática Industrial (ITMATI).

## ABSTRACT

Standard survival models are used to analyse time-to-event data. They assume that all individuals would experience the event of interest if they could be followed-up long enough. In many situations there are subjects, known as *cured*, that will never experience that event. To incorporate this cure fraction, classical survival analysis has been extended to cure models. In particular, mixture cure models allow to estimate the probability of cure and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates. We fill this important gap by considering both vector and functional covariates and proposing a single-index model for dimension reduction. The methodology will be applied to a cardiotoxicity dataset from the University Hospital of A Coruña (CHUAC).

**Keywords**: cardiotoxicity; censored data; kernel estimator; survival analysis.

## REFERENCES

Boag, J. W. (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society Series B, 11, 15–53.

Bouaziz, O. and Lopez, O. (2010) Conditional density estimation in a censored single-index model. Bernoulli, 16, 514–542.

Farewell, V. T. (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 38, 1041–1046.

Kuk, A. Y. and Chen, C. H. (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika, 79, 531–541.

Laska, E. M. and Meisner, M. J. (1992) Nonparametric estimation and testing in a cure model. Biometrics, 48, 1223–1234.

Li, C. and Taylor, J. M. G. (2000) A semi-parametric accelerated failure time cure model. Statistics in Medicine, 21, 3235–3247.

López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017a) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Computational Statistics & Data Analysis, 105, 144–165.

López-Cheda, A., Jácome, M. A. and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. TEST, 26, 353–376.

López-Cheda, A., Jácome, M. A. and López-de-Ullibarri, I. (2021) npcure: An R Package for Nonparametric Inference in Mixture Cure Models. R Journal, 13(1), 21–41.

Maller, R. A. and Zhou, S. (1992) Estimating the proportion of immunes in a censored sample. Biometrika, 79, 731–739.

Mazzotta, M., Krasniqi, E., Barchiesi, G., Pizzuti, L., Tomao, F., Barba, M. and Vici, P. (2019) Long-term safety and real-world effectiveness of trastuzumab in breast cancer. Journal of Clinical Medicine, 8(2), 254.

Patilea, V. and Van Keilegom, I. (2020) A general approach for cure models in survival analysis. Annals of Statistics, 48(4), 2323–2346.

Peng, Y. and Dear, K. B. (2000) A nonparametric mixture model for cure rate estimation. Biometrics, 56, 237–243.

Strzalkowska-Kominiak, E. and Cao, R. (2013) Maximum likelihood estimation for conditional distribution single-index models under censoring. Journal of Multivariate Analysis, 114, 74–98.

Sy, J. P. and Taylor, J. M. G. (2000) Estimation in a Cox proportional hazards cure model. Biometrics, 56, 227–236.

Wadhwa, D., Fallah-Rad, N., Grenier, D. et al. (2009) Trastuzumab mediated cardiotoxicity in the setting of adjuvant chemotherapy for breast cancer: a retrospective study. Breast Cancer Research and Treatment, 117, 357–364.

Xu, J. and Peng, Y. (2014) Nonparametric cure rate estimation with covariates. Biometrics, 42, 1–17.

Yamaguchi, K. (1992) Accelerated failure-time regression model with a regression model of surviving fraction: an analysis of permanent employment in Japan. Journal of the American Statistical Association, 87, 284–292.

# Premios modalidade B

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ANALIZANDO AS ESTRATEXIAS DE ESCAPE EN LARVAS DE PEIXE CEBRA MEDIANTE REGRESIÓN MULTIMODAL CIRCULAR

María Alonso-Pena[1] e Rosa M. Crujeiras[1]

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

## RESUMO

Analizar a dirección de escape de animais en función de covariables é un problema que require de técnicas estatísticas alén dos métodos de regresión clásicos. Ademais da periodicidade do ángulo de escape, que esixe a utilización de métodos para variables circulares, os datos de escape animal adoitan requirir da exploración das direccións preferentes, en lugar da dirección media ou esperada. Neste traballo propoñemos o uso dun método non paramétrico para estimar as modas condicionais locais nun experimento con peixes cebra, dende a perspectiva dos modelos de regresión. Presentaremos os algoritmos de estimación e investigaremos o comportamento asintótico do estimador, así como o seu funcionamento con datos simulados. A nova metodoloxía é empleada para modelar o comportamento de escape dun grupo de larvas de peixe cebra cando fuxen dun depredador robot. De xeito máis xeral, o enfoque proposto neste traballo pode aplicarse a moitos outros problemas relativos ao comportamento animal ou a outros campos.

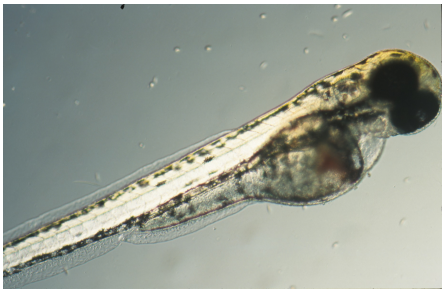
**Palabras e frases chave**: Regresión multimodal; Datos circulares; Estimación tipo núcleo; Escapoloxía animal; Regresión non paramétrica

## 1. INTRODUCIÓN

Existe unha ampla literatura no campo da bioloxía que trata o estudo da elección de orientación ou comportamento de escape en animais, e a influencia de covariables na resposta de escape (Scapini et al., 2002; Marchetti e Scapini, 2003; Card e Dickinson, 2008; Obleser et al., 2016; Sato et al., 2019). Un caso interesante é a análise das direccións de escape en animais cando fuxen dos seus depredadores e, aínda que a lóxica suxire que os animais deberían fuxir en dirección oposta aos seus depredadores, este non é sempre o caso, e cómpre analizar as respostas de escape. Máis especificamente, estamos interesados nas direccións de escape de larvas de peixes cebra (*Danio rerio*), como as que se mostran no panel esquedo da Figura 1, e como o ángulo no que o depredador se aproxima aos peixes afecta ao comportamento á hora de escapar. Os datos, obtidos de Nair et al. (2017a), conteñen as direccións de escape dun grupo de larvas de peixe cebra retidos nun acuario e os ángulos nos que un robot imitando un depredador se aproximaron a cada peixe. O panel dereito da Figura 1 mostra un esquema do experimento.

Cando se trata de analizar este tipo de datos con métodos de regresión clásicos, podemos atoparnos dous problemas: i) a periodicidade da variable determinando a dirección de escape, que é, por definición, unha variable circular e ii) a necesidade de estimar as direccións máis verosímiles, condicionadas a distintas covariables, en lugar da dirección media ou esperada (condicional).

A teoría clásica sobre datos circulares (observacións definidas sobre a circunferencia unidade) está presente dende fai décadas (Fisher, 1993; Mardia e Jupp, 2000; Jammalamadaka e SenGupta, 2001), pero o seu uso na práctica foi limitado debido á falta de observacións circulares precisas. Os avances tecnolóxicos fixeron posible o rexistro preciso deste tipo de datos, incrementando o interés no campo da estatística circular nos últimos anos (Pewsey et al., 2013; Ley e Verdebout, 2017). Ademais de en bioloxía, podemos atopar datos circulares en moitos outros ámbitos: xeoloxía (SenGupta e Rao, 1966), ciencias ambientais e oceanografía (Oliveira et al., 2013), medicina



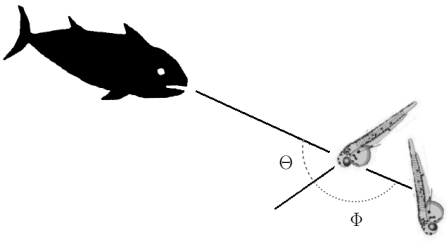



Figura 1: Esquerda: fotografía dunha larva de peixe cebra de Wikimedia Commons (2008). Dereita: esquema do experimento, onde $\Theta$ indica a dirección na que se aproxima o depredador e $\Phi$ a dirección de escape (elaboración propia a partir da imaxe de Wikimedia Commons (2014)).

(Mooney et al. 2003) ou ecoloxía (Ameijeiras-Alonso et al., 2019). A peculiar natureza deste tipo de observacións pon de manifesto a necesidade de crear ferramentas inferenciais específicas máis aló de aquelas pensadas para datos na recta real.

Como no caso que nos ocupa, no que temos outra variable influíndo na dirección de escape, podemos estudar os datos circulares dende a perspectiva da regresión. Dependendo da natureza da covariable, podemos distinguir dous escenarios distintos: se a covariable é escalar, podemos representar a curva de regresión na superficie dun cilindro, no que a altura do cilindro indica a magnitude da covariable e o ángulo representa o valor da resposta circular, como se representa nos paneis superiores da Figura 2. Por outra banda, se a natureza da covariable é tamén circular, a curva de regresión pódese representar na superficie dun toro, como se mostra nos paneis inferiores da Figura 2. Distintas propostas de modelos de regresión paramétricos involucrando variables circulares poden atoparse en Jammalamadaka e SenGupta (2001). Porén, estes modelos poden non ser o suficientemente flexibles para modelar relacións máis complexas entre as variables, polo que os modelos non paramétricos preséntanse como unha boa alternativa. Di Marzio et al. (2012) propuxeron un método tipo núcleo para regresión con resposta circular, no que a estimación vén dada polo suavizado das compoñentes seno e coseno da resposta.

Os métodos anteriormente citados consideran a media condicional como a función a estimar. Non obstante, como se expuxo no punto ii), o enfoque clásico de *regresión á media* pode non ser axeitado en casos onde a densidade condicional é multimodal. A Figura 2 presenta datos simulados de modelos de regresión con resposta circular nos que a densidade condicional da resposta sobre a explicativa é bimodal. Nos casos correspondentes aos paneis esquerdos da Figura 2, a media condicional ou función de regresión *usual* non está nin sequera definida, dado que as densidades condicionais consideradas son bimodais e simétricas, non estando a media circular definida neste caso. No seu lugar, as modas condicionais locais preséntanse como unha mellor alternativa para modelar a relación entre as variables, resumindo os valores condicionais *máis probables* en lugar da esperanza condicional. Esta idea lévanos a chamada *regresión multimodal*, na que no lugar dunha función, o obxectivo é estimar unha multifunción ou función de avaliación múltiple.

A idea de estimar as modas locais da densidade condicional no contexto euclídeo foi primeiramente introducida por Scott (1992). Einbeck e Tutz (2006) propuxeron a versión condicional do algoritmo mean shift (Fukunaga e Hostetler, 1975; Cheng, 1995; Comaniciu e Meer, 2002) para estimar a multifunción de regresión baseándose nun estimador tipo núcleo da densidade condicional. As propiedades teóricas deste estimador foron estudadas por Chen et al. (2016) e estes modelos foron estendidos ao contexto de datos con erros de medición por Zhou e Huang (2016). Unha revisión recente da regresión multimodal con variables escalares pode atoparse en Chen (2018). O algoritmo mean shift foi xeneralizado ao caso de variables direccionais por Oba et al. (2005) no contexto do clustering non paramétrico, e foi tamén estudado por Kobayashi e Otsu (2010) nese mesmo contexto. A converxencia do algoritmo e outros resultados teóricos foron obtidos recentemente por Zhang e Chen (2020).

O obxectivo deste traballo é introducir o método da regresión multimodal non paramétrica

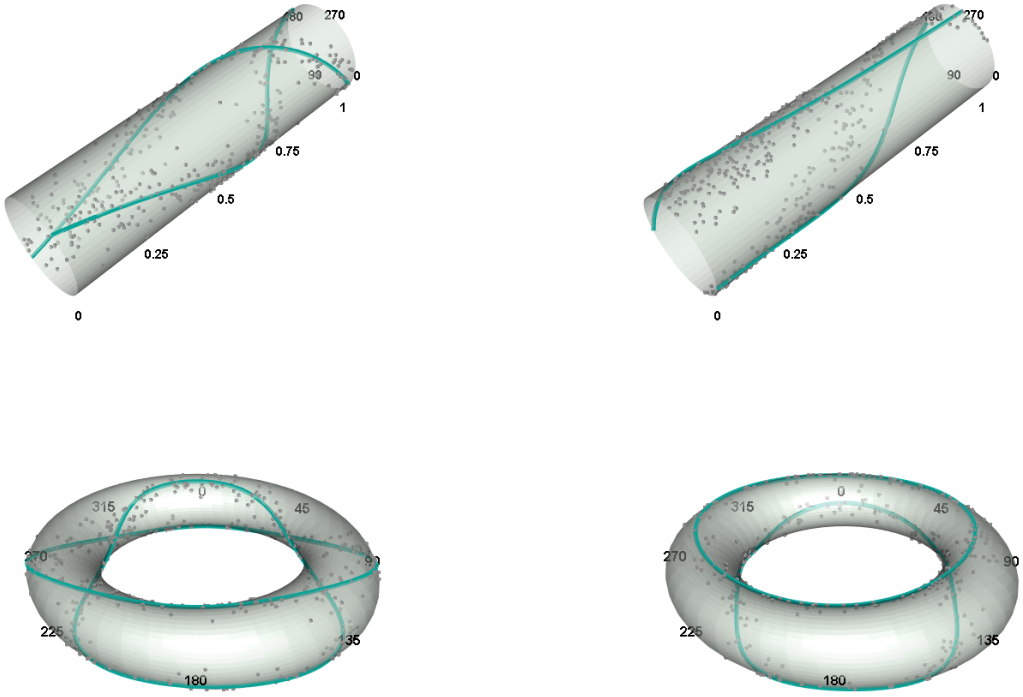



Figura 2: Representacións no cilindro e no toro de datos simulados e as verdadeiras multifuncións de regresión para os modelos models LC-1 (arriba e esquerda), LC-2 (arriba e dereita) con concentración $\tau = 8$; e modelos CC-1 abaixo e esquerda) e CC-2 (abaixo e dereita) con concentración $\tau = 12$. Os tamaños mostrais son $(n_1, n_2) = (200, 200)$ para todos os modelos.

circular para analizar situacións onde as modas locais condicionais son un mellor representante da relación entre as variables que a función de regresión clásica. A nosa meta é empregar esta técnica para estudar como a dirección de escape dos peixes cebra está influenciada polo ángulo no que se aproxima o depredador.

A estrutura do documento é a que segue: na Sección 2 presentamos o escenario da regresión multimodal circular e detallamos os algoritmos de estimación. Algúns resultados teóricos enúncianse na Sección 3, mentres que o problema da selección dos parámetros de suavizado estúdase na Sección 4. A Sección 5 mostra un estudo de simulación no que se analiza o comportamento dos estimadores na práctica. Finalmente, a Sección 6 contén a análise dos datos dos peixes cebra coa metodoloxía proposta.

## 2. O USO DO MEAN SHIFT CIRCULAR NA REGRESIÓN MULTIMODAL

O obxectivo da regresión multimodal non paramétrica é estimar a densidade condicional da variable resposta sobre a variable explicativa e, despois, computar as modas locais condicionais co algoritmo mean shift (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu e Meer, 2002). Nesta sección detallaremos o algoritmo de estimación deseñado para estimar a multifunción de regresión multimodal no contexto onde a variable resposta é de natureza circular e onde a variable explicativa pode ser tanto circular como escalar.

Considérese unha variable resposta circular, $\Phi$, con soporte na circunferencia unidade, $\mathbb{T} = (-\pi, \pi]$ e unha variable explicativa que pode ser tanto unha variable escalar $X$ con soporte en

$\Omega \subset \mathbb{R}$ ou unha explicativa circular con soporte en $\mathbb{T}$. Denotaremos por $\Delta$ unha variable explicativa xenérica, con soporte na recta real ou na circunferencia unidade. Sexa $\{(\Delta_j, \Phi_j)\}_{j=1}^n$ unha mostra aleatoria de $(\Delta, \Phi)$. Para modelar a relación entre a explicativa e a resposta, consideramos a multifunción de regresión modal (Einbeck e Tutz, 2006) que, para cada $\delta$ está definida como o conxunto de modas locais (ou máximos locais) da función de densidade condicional:

$$M(\delta) = \left\{ \phi : \frac{\partial}{\partial \phi} f(\phi|\delta) = 0, \quad \frac{\partial^2}{\partial \phi^2} f(\phi|\delta) < 0 \right\}, \tag{1}$$

onde $f(\phi|\delta)$ é a densidade condicional de $\Phi$ dado o valor de $\Delta$. A estimación de $M$ lévase a cabo mediante un enfoque indirecto: primeiramente, estimamos a densidade condicional e, despois, calcúlanse as modas locais condicionaiss. Para a estimación da densidade condicional $f(\phi|\delta)$ utilizaremos un estimador tipo núcleo (Di Marzio et al, 2016). Se a variable explicativa é escalar ($\Delta = X$), o estimador vén dado por

$$\hat{f}(\phi|x) = \frac{\sum_{j=1}^n L_h\left(x - X_j\right) K_\kappa\left(\phi - \Phi_j\right)}{\sum_{j=1}^n L_h\left(x - X_j\right)}.$$

Neste caso $L_h(\cdot)$ é unha función núcleo *linear* ou usual, con ancho de banda $h$ e $K_\kappa(\cdot)$ é unha función núcleo circular con parámetro de concentración $\kappa$. Se a explicativa é circular ($\Delta = \Theta$), estimaremos a densidade condicional como

$$\hat{f}(\phi|\theta) = \frac{\sum_{j=1}^n K_\nu\left(\theta - \Theta_j\right) K_\kappa\left(\phi - \Phi_j\right)}{\sum_{j=1}^n K_\nu\left(\theta - \Theta_j\right)},$$

onde o núcleo asociado á explicativa ten $\nu$ como parámetro de concentración e o núcleo asociado a $\Phi$ ten concentración $\kappa$. De agora en adiante denotaremos os pesos correspondentes á variable explicativa ($X$ ou $\Theta$) no punto $\delta$ como $w_\delta(\Delta_j)$, $j = 1, ..., n$. Nótese que estes pesos dependen dun parámetro de concentración $h$ ou $\nu$ (dependendo da natureza de $\Delta$). Así,

$$\hat{f}(\phi|\delta) = \frac{1}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K_\kappa(\phi - \Phi_j), \quad \hat{f}(\delta) = \frac{1}{n} \sum_{j=1}^n w_\delta(\Delta_j).$$

En consecuencia, o estimador da multifunción de regresión modal (1) vén dado por

$$\widehat{M}(\delta) = \left\{ \phi : \frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = 0, \quad \frac{\partial^2}{\partial \phi^2} \hat{f}(\phi|\delta) < 0 \right\}. \tag{2}$$

Asumiremos que a función núcleo circular asociada á variable resposta satisfai

$$K_\kappa(\cdot) = c_\kappa K[\kappa(1 - \cos(\cdot))], \tag{3}$$

onde $c_\kappa$ é unha constante normalizadora que depende de $\kappa$. A densidade de von Mises é un exemplo de núcleo circular satisfacendo esta condición. Para obter os máximos locais de $\hat{f}(\delta|\phi)$, establecemos a condición necesaria de punto crítico: $\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = 0$. Polo tanto, se aplicamos (3), temos que

$$\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = \frac{\kappa c_\kappa}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K'[\kappa(1 - \cos(\phi - \Phi_j))] \sin(\phi - \Phi_j). \tag{4}$$

Por conseguinte, a derivada do estimador da densidade condicional con respecto a $\phi$ é unha suma poderada dos senos das diferencias de cada observación ao punto $\phi$. Neste caso, a función seno utilízase para medir a variación ou diferenzas entre as observacións e o punto $\phi$. Isto é bastante intuitivo dado que se $\phi = \Phi_j$, entón $\sin(\phi - \Phi_j) = 0$. Ademais, se a diferenza $\phi - \Phi_j$ é moi pequena, entón $\sin(\phi - \Phi_j) \approx \phi - \Phi_j$. Expandindo o último factor no lado dereito de (4), obtemos

$$\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = \frac{\kappa c_\kappa}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K'[\kappa(1 - \cos(\phi - \Phi_j))](\sin\phi\cos\Phi_j - \cos\phi\sin\Phi_j),$$

e igualándoo a cero, temos

$$\sin\phi\sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\cos\Phi_j = \cos\phi\sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\sin\Phi_j,$$

onde $T(\cdot) = c_T K'[\kappa(1-\cos(\cdot))]$. Polo tanto, se denotamos

$$S_\delta(\phi) = \sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\sin\Phi_j \quad \text{e} \quad C_\delta(\phi) = \sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\cos\Phi_j,$$

temos que se $S_\delta(\phi) \neq 0$ ou $C_\delta(\phi) \neq 0$, entón $\phi = \text{atan2}\left(S_\delta(\phi), C_\delta(\phi)\right)$, onde o operador $\text{atan2}(a,b)$ devolve o ángulo entre o eixo das $x$ e o vector que vai da orixe a $(b,a)$ (see Jammalamadaka e SenGupta, 2001, Capítulo 1). Deste xeito, obtemos que o estimador modal $\phi_m \equiv \phi_m(\delta)$ vén dado por

$$\phi_m = \tilde{\omega}(\phi_m) = \text{atan2}\left(S_\delta(\phi_m), C_\delta(\phi_m)\right).$$

Nótese que a función $\tilde{\omega}(\phi)$ devolve unha media circular ponderada das observacións, dado que $S_\delta(\phi)$ é unha suma ponderada de $\sin\Phi_j$ e $C_\delta(\phi)$ é unha suma ponderada de $\cos\Phi_j$ (onde os pesos dependen do punto $(\delta,\phi)$). Como na anterior expresión temos unha ecuación de punto fixo, utilizamos un algoritmo tipo mean shift para obter o estimador da moda condicional. Definimos a función mean shift circular como

$$\tilde{m}(\phi) = \sin(\tilde{\omega}(\phi) - \phi).$$

Como se comentou anteriormente, dado que para valores pequenos de $\tilde{\omega}(\phi)-\phi$ temos $\sin(\tilde{\omega}(\phi)-\phi) \approx \tilde{\omega}(\phi)-\phi$, a función seno utilízase para medir a variación de $\phi$ a $\tilde{\omega}(\phi)$. Ademais, para a moda local do estimador da densidade condicional, a función mean shift circular toma o valor cero. En consecuencia, a multifunción de regresión estimada (2) obtense utilizando o procedemento mean shift circular, que se describe no Algoritmo 1. Nótese que o número de puntos iniciais, $p$, pode ser diferente para cada valor de $\delta$, e para inicializar o algoritmo en rexións próximas aos datos, recomendamos unha inicialización local onde, para cada valor $\delta$, os valores iniciais son os cuartís circulares (véxase Fisher, 1993) das observacións da varriable resposta máis próximas a $\delta$.

**Algoritmo 1:** Mean shift condicional circular

**Datos:** Mostra $\{(\Delta_i,\Phi_i\}_{i=1}^n$, parámetros de suavizado $\kappa$ e $h/\nu$.

1. Inicializar puntos da malla $\mathcal{S} \subset \Omega$ se $\Delta = X$ ou $\mathcal{T} \subset \mathbb{T}$ se $\Delta = \Theta$.
2. Para cada $\delta \in \mathcal{S}$ (ou $\delta \in \mathcal{T}$), seleccionar valores iniciales $\phi_0^{(1)}(\delta), ..., \phi_0^{(p)}(\delta)$.
3. Para $k = 1, ..., p$ iterar ata alcanzar converxencia:

$$\phi_{l+1}^{(k)} = \text{atan2}\left(\sum_{i=1}^{n} w_\delta(\Delta_i)T(\phi_l^{(k)}-\Phi_i)\sin\Phi_i, \sum_{i=1}^{n} w_\delta(\Delta_i)T(\phi_l^{(k)}-\Phi_i)\cos\Phi_i\right),$$

con $l = 0, 1, ...$

## 3. ALGUNHAS CONSIDERACIÓNS TEÓRICAS

O obxectivo desta sección é dar taxas de converxencia asintótica do estimador non paramétrico para a regresión multimodal introducido na Sección 2. Nótese que as métricas de erro usuais en regresión tipo núcleo (como o Erro Cadrático Medio Integrado ou o Erro Cadrático Integrado) non son axeitadas para medir a calidade do estimador (2) dado que, no contexto da regresión multimodal, o noso obxectivo é estimar a multifunción (1) e, polo tanto, para cada valor da variable explicativa hai, posiblemente, un conxunto de valores da variable resposta. En consecuencia, seguiremos o traballo de Chen et al. (2016) e consideraremos erros puntuais e globais baseados nunha distancia entre conxuntos.

En primeiro lugar, presentamos a distancia de Hausdorff que, para dous conxuntos $A, B \subset \mathbb{R}^q$ defínese como

$$\text{Haus}(A,B) = \max\left\{\sup_{x\in A} d(x,B), \sup_{x\in B} d(x,A)\right\}, \tag{5}$$

onde $d(x, A) = \inf_{z\in A} ||x - z||$. Esta distancia mide como de próximos están dous conxuntos definidos en espazos euclídeos. Na regresión multimodal para variables escalares, a distancia de Hausdorff utilízase para medir a distancia da verdadeira multifunción á súa versión estimada (para un punto fixado da variable explcativa). Porén, no caso que nos ocupa, $M(\delta)$ e $\widehat{M}(\delta)$ son subconxuntos de $\mathbb{T}$ e, polo tanto, é necesario xeneralizar esta distancia considerando, para $A, B \subset \mathbb{T}$,

$$\widetilde{\text{Haus}}(A, B) = \max\left\{ \sup_{x\in A} \tilde{d}(x, B), \sup_{x\in B} \tilde{d}(x, A) \right\}, \tag{6}$$

con $\tilde{d}(x, A) = \inf_{z\in A} 1 - \cos(x - z)$. Utilizaremos a distancia definida en (6) para construír unha medida de erro puntual na regresión multimodal circular. Definimos o erro puntual como

$$\tilde{\Lambda}(\delta) = \widetilde{\text{Haus}}(M(\delta), \widehat{M}(\delta)). \tag{7}$$

O erro puntual mide como de próximas están a multifunción de regresión verdadeira e a estimada para cada posible valor $\delta$ da variable explicativa. Para obter unha medida de erro global, introducimos o Erro Medio Circular Integrado modal ($\text{EMCI}_m$), definido como

$$\text{EMCI}_m(\widehat{M}) = \mathbb{E}\left[ \int_{\delta\in\text{Supp}(\Delta)} \tilde{\Lambda}(\delta) d\delta \right],$$

onde $\text{Supp}(\Delta)$ denota o soporte de $\Delta$ que, recordemos pode denotar tanto unha variable escalar como unha variable circular. Esta medida é o valor esperado do erro puntual integrado, e está baseado na versión integrada do chamado *Erro Cadrático Medio Circular*, introducido po Kim e SenGupta (2017) como o análogo circular do Erro Cadrático Medio en variables escalares. As seguintes proposicións mostran a consistencia dos estimadores.

**Proposición 1.** *Sexa $X$ unha variable explicativa escalar e $\Phi$ unha variable resposta circular. Considérese o estimador da regresión multimodal en (2). Baixo certas condicións de regularidade, tense*

$$\tilde{\Lambda}(x) = O(h^2 + \kappa^{-1}) + O_P\left(\sqrt{\frac{\kappa^{3/2}}{nh}}\right)$$

*e*

$$EMCI_m(\widehat{M}) = O(h^2 + \kappa^{-1}) + O\left(\sqrt{\frac{\kappa^{3/2}}{nh}}\right),$$

*cando $\kappa \to \infty$, $h \to 0$ e $nh\kappa^{-(1+2r)/2} (\log n)^{-1} \to \infty$.*

**Proposición 2.** *Sexa $\Theta$ unha variable explicativa circular e $\Phi$ unha variable resposta circular. Considérese o estimador da regresión multimodal en (2). Baixo certas condicións de regularidade, tense*

$$\tilde{\Lambda}(\theta) = O(\nu^{-1} + \kappa^{-1}) + O_P\left(\sqrt{\frac{\kappa^{3/2}\nu^{1/2}}{n}}\right)$$

*e*

$$EMCI_m(\widehat{M}) = O(\nu^{-1} + \kappa^{-1}) + O\left(\sqrt{\frac{\kappa^{3/2}\nu^{1/2}}{n}}\right),$$

*cando $\kappa \to \infty$, $\nu \to \infty$ e $n\kappa^{-(1+2r)/2}\nu^{-1/2} (\log n)^{-1} \to \infty$.*

Os resultados previos mostran que os erro puntuais $\tilde{\Delta}(x)$ e $\tilde{\Delta}(\theta)$ converxen a cero coa mesma taxa de converxencia que a primeira derivada parcial, con respecto a variable resposta, dos estimadores tipo núcleo da densidade conxunta, $\frac{\partial}{\partial\phi}\hat{f}(x, \phi)$ e $\frac{\partial}{\partial\phi}\hat{f}(\theta, \phi)$.

## 4. A SELECCIÓN DOS PARÁMETROS DE SUAVIZADO

Como adoita suceder en estimación tipo núcleo, a selección dos parámetros de suavizado resulta crucial en regresión multimodal. No contexto da regresión tipo núcleo clásica na recta real (Fan e Gijbels, 1996), un valor alto da fiestra $h$ produce un estimador sobresuvizado, mentres que un valor pequeno de $h$ implica unha estimación infrasuavizada da función de regresión. Pola contra, o comportamento do parámetro de concentración presente na regresión tipo núcleo circular (Di Marzio et al. 2009, 2012) é inverso: cando a concentración $\kappa$ e grande, obtemos un estimador infrasuavizado, mentres que un valor pequeno de $\kappa$ produce unha versión sobresuavizada do estimador.

Así e todo, no contexto da regresión multimodal son necesarios dous parámetros de suavizado: un asociado á variable explicativa e outro asociado á variable resposta. O rol que desempeñan estes parámetros na estimación da multifunción de regresión é moi diferente. O parámetro asociado á variable explicativa controla o suavizado do estimador da multifunción, xogando un papel similar ao do parámetro de suavizado na regresión tipo núcleo clásica. Pola contra, o parámetro asociado á variable resposta inflúe no número de modas estimadas. A razón detrás deste comportamento é que un estimador infrasuavizado da densidade condicional dará lugar a moitas modas locais estimadas, producindo un alto número de *ramas* estimadas da multifunción.

A literatura estatística sobre a selección dos parámetros de suavizado para regresión multimodal no contexto de variables escalares é relativamente escasa. Dado que a estimación se leva a cabo mediante a obtención dos máximos locais da densidade condicional, Einbeck e Tutz (2006) recomendan utilizar métodos deseñados para a estimación da densidade condicional. Non obstante, estes métodos poden non ser idóneos na práctica, dado que a estimación da moda está relacionada pero non é equivalente á estimación da densidade. Como apuntan Casa et al. (2020), unha estimación da densidade pode estar próxima á verdadeira densidade en termos do Erro Cadrático Integrádo, pero ter moitas modas esrimadas que, no caso da regresión multimodal, daría lugar a moitas ramas na multifunción estimada. Zhou e Huang (2019) propuxeron dous métodos diferentes para obter parámetros de suavizado na práctica no contexto da regresión multimodal con variables escalares. O primeiro, coñecido como validación cruzada modal, aspira a equilibrar o número de modas locais estimados e a distancia da multifunción estimada aos datos. Aínda que este método mostra un bo comportamento na práctica, nada asegura que minimizar a función de validación cruzada modal minimizará o Erro Cadrático Medio Integrado modal do estimador ou calquera outro criterio de erro. O segundo procedemento proposto por Zhou e Huang (2019) é minimizar o Erro Cadrático Integrado modal do estimador (a versión integrada da distancia de Hausdorff entre o estimador e a verdadeira multifunción) utilizando un método de remostraxe baseado nunha mestura de regresións parámetricas. Outro criterio, proposto por Chen et al. (2016), consiste en construír unha banda de predicción para a multifunción de regresión e posteriormente seleccionar os parámetros que minimicen unha función de perda definida como o volume de dita banda. Porén, os autores asumen que o parámetro de suavizado é o mesmo para ambas variables, o cal non está xustificado, especialmente tendo en conta o diferente papel que xogan os dous parámetros. Ademais, a selección do parámetro depende do nivel de predicción previamente fixado.

**Validación cruzada modal para regresión circular.** Unha das vantaxes da validación cruzada modal de Zhou e Huang (2019) é que a súa adaptación a casos máis complexos como o da regresión circular é case inmediata. Aínda que as propiedades teóricas deste procedemento non están ben estudadas, o comportamento deste método na práctica resulta satisfactorio. Para o escenario presentado na Sección 2, onde a variable resposta é circular, a validación cruzada modal consiste en seleccionar os parámetros $g$ e $\kappa$ (onde $g$ representa tanto $h$ ou $\nu$, dependendo da natureza da variable explicativa) mediante a minimización de

$$CV(g,\kappa) = \frac{1}{n}\sum_{i=1}^{n} \tilde{d}(\widehat{M}_{-i}^{g,\kappa}(\Delta_i), Y_i) N_{-i}(\Delta_i), \tag{8}$$

onde $\tilde{d}(x, A) = \inf_{z\in A} 1 - \cos(x - z)$, $\widehat{M}_{-i}^{g,\kappa}$ é o estimador de $M$ utilizando os datos $\{(\Delta_j, \Phi_j) : j \neq i\}$ construído cos parámetros $g$ e $\kappa$ e $N_{-i}(\Delta_i)$ denota o número de modas locais estimadas cando $\Delta = \Delta_i$. Este método non se basea en fundamentacións teóricas e precísanse de ensaios computacionais para avaliar a súa eficacia.

## 5. ESTUDO DE SIMULACIÓN

Nesta sección analizamos o comportamento do estimador tanto no caso no que a variable explicativa é escalar como no caso onde está presenta unha natureza circular. En primeiro lugar presentamos os escenarios de simulación e de seguido, amosamos os resultados obtidos.

**Escenarios de simulación.** Nos nosos exemplos simulados, a mostra está dividida en dous grupos, con cada grupo correspondéndose cunha rama da multifunción obxectivo. Para cada observación utilizamos o subíndice $ji$, onde $j$ denota o grupo ou número de rama e $i$ denota o número de observación dentro de cada grupo. Ademais, $n_j$ denota o tamaño mostral para o $j$-ésimo grupo. Nótese que estes grupos subxacentes non son coñecidos na práctica e non dispoñemos de información sobre eles. Os modelos simulados móstranse na Tabla 1.

O primeiro modelo en cada escenario correspóndese con dúas curvas paralelas ou, visto dende outra perspectiva, a únha curva de regresión cun erro bimodal. No que se refire ao segundo modelo de cada escenario, as dúas curvas non son paralelas. En todos os casos, os tamaños mostrais son $(n_1, n_2) \in \{(100,100),(100,200),(200,200),(200,300),(300,300)\}$. Exemplos de datos simulados de todos os modelos poden verse na Figura 2 xunto coas multifuncións verdadeiras.

| Modelo | Xeneración da mostra | Multifunción de regresión |
|---|---|---|
| LC-1 | $\Phi_{1i} = (6\tan^{-1}(2.5X_{1i} - 3) + \varepsilon_{1i})(\text{mod } 2\pi)$<br>$\Phi_{2i} = (\pi + 6\tan^{-1}(2.5X_{1i} - 3) + \varepsilon_{1i})(\text{mod } 2\pi)$<br>$X_1, X_2 \sim U(0,1)$ | $M(x) = \{6\tan^{-1}(2.5x - 3),$<br>$\pi + 6\tan^{-1}(2.5x - 3)\}(\text{mod } 2\pi)$ |
| LC-2 | $\Phi_{1i} = (\text{atan2}(\sin(3X_{1i}^2), \cos(3X_{1i}^2)) + \varepsilon_{1i})(\text{mod } 2\pi)$<br>$\Phi_{2i} = (\pi/2 + 2\tan^{-1}(10X_{2i} - 1/2) + \varepsilon_{2i})(\text{mod } 2\pi)$<br>$X_1, X_2 \sim U(0,1)$ | $M(x) = \{\text{atan2}(\sin(3x^2), \cos(3x^2)),$<br>$\pi/2 + 2\tan^{-1}(10x + 1/2)\}(\text{mod } 2\pi)$ |
| CC-1 | $\Phi_{1i} = (2\cos\Theta_{1i} + \varepsilon_{1i})(\text{mod } 2\pi)$<br>$\Phi_{2i} = (3\pi/4 + 2\cos\Theta_{1i} + \varepsilon_{2i})(\text{mod } 2\pi)$<br>$\Theta_1, \Theta_2 \sim$ Circular Uniform | $M(\theta) = \{2\cos\theta, 3\pi/4 + 2\cos\theta\}(\text{mod } 2\pi)$ |
| CC-2 | $\Phi_{1i} = (3/4\cos\Theta_{1i} - \pi/2 + \varepsilon_{1i})(\text{mod } 2\pi)$<br>$\Phi_{2i} = (\pi/2 - \cos\Theta_{1i} + \varepsilon_{2i})(\text{mod } 2\pi)$<br>$\Theta_1, \Theta_2 \sim$ Circular Uniform | $M(\theta) = \{3/4\cos\theta - \pi/2, \pi/2 - 2\cos\theta\}(\text{mod } 2\pi)$ |

Táboa 1: Modelos simulados. LC denota explicativa escalar e resposta circular e CC denota explicativa circular e resposta circular. (mod $2\pi$) denota módulo $2\pi$. En todos os modelos $\varepsilon_{1i}, \varepsilon_{2i} \sim vM(0,\tau)$ with $\tau \in \{6,8,10\}$.

Para medir o comportamento dos estimadores, o Erro Medio Circular Integrado modal ($\text{EMCI}_m$) foi aproximado tras xerar 500 réplicas Monte Carlo de datos simulados e calculando a media do Erro Integrado Circular modal ($\text{EIC}_m$) do estimador multimodal:

$$\text{EIC}_m(\widehat{M}) = \int_{\delta \in \text{Supp}(\Delta)} \widetilde{\Lambda}(\delta) d\delta,$$

onde as integrais foron aproximadas numericamente mediante a regra de Simpson e $\widetilde{\Lambda}$ está definida en (7). Nos experimentos de simulación, os parámetros de suavizado foron seleccionados mediante validación cruzada modal. Os resultados compáranse cos obtidos tras utilizar os parámetros que minimizan o $\text{EIC}_m$, que se toman como referencia.

**Resultados.** O $\text{EMCI}_m$ estimado para todos os modelos pode atoparse na Táboa 2. Como se esperaba, o valor estimado do $\text{EMCI}_m$ xeralmente diminúe ao incrementar o tamaño mostral. Os poucos casos nos que non se observa este comportamento son cando os tamaños mostrais en cada grupo non son iguais, é dicir$(n_1, n_2) \in \{(100,200),(200,300)\}$. Ademais, un valor grande da concentración do erro tamén da lugar a unha menor estimación do $\text{EMCI}_m$. O desempeño do criterio de validación cruzada modal é tamén máis que aceptable, dado que para valores grandes do tamaño mostral, os valores estimados do $\text{EMCI}_m$ cando os parámetros de suavizado se seleccionan mediante este criterio están moi preto dos valores obtidos cos parámetros *óptimos*. O peor rendemento

obtense co modelo LC-2 con $\kappa = 6$. Como se mostra no panel de arriba e esquerda da Figura 2, cando $x$ toma valores próximos a 1, as dúas ramas da multifunción están moi próximas, o que fai difícil discernir entre os dous grupos cando a concentración do erro é baixa.

| | | $\tau = 6$ | | $\tau = 8$ | | $\tau = 10$ | |
|---|---|---|---|---|---|---|---|
| Model | $(n_1, n_2)$ | B | CV | B | CV | B | CV |
| LC-1 | $(100, 100)$ | 0.021 | 0.034 | 0.016 | 0.023 | 0.013 | 0.018 |
| | $(100, 200)$ | 0.016 | 0.039 | 0.012 | 0.024 | 0.011 | 0.018 |
| | $(200, 200)$ | 0.011 | 0.019 | 0.008 | 0.012 | 0.007 | 0.010 |
| | $(200, 300)$ | 0.010 | 0.016 | 0.007 | 0.011 | 0.006 | 0.009 |
| | $(300, 300)$ | 0.007 | 0.012 | 0.006 | 0.008 | 0.005 | 0.006 |
| LC-2 | $(100, 100)$ | 0.016 | 0.035 | 0.012 | 0.025 | 0.010 | 0.018 |
| | $(100, 200)$ | 0.016 | 0.070 | 0.010 | 0.039 | 0.008 | 0.024 |
| | $(200, 200)$ | 0.010 | 0.023 | 0.007 | 0.012 | 0.005 | 0.009 |
| | $(200, 300)$ | 0.009 | 0.038 | 0.006 | 0.017 | 0.005 | 0.011 |
| | $(300, 300)$ | 0.007 | 0.015 | 0.005 | 0.009 | 0.004 | 0.007 |
| CC-1 | $(100, 100)$ | 0.144 | 0.395 | 0.117 | 0.213 | 0.102 | 0.148 |
| | $(100, 200)$ | 0.120 | 0.182 | 0.097 | 0.147 | 0.087 | 0.120 |
| | $(200, 200)$ | 0.066 | 0.089 | 0.054 | 0.066 | 0.047 | 0.057 |
| | $(200, 300)$ | 0.058 | 0.068 | 0.046 | 0.054 | 0.040 | 0.045 |
| | $(300, 300)$ | 0.042 | 0.056 | 0.035 | 0.045 | 0.030 | 0.038 |
| CC-2 | $(100, 100)$ | 0.116 | 0.227 | 0.092 | 0.173 | 0.080 | 0.150 |
| | $(100, 200)$ | 0.085 | 0.232 | 0.068 | 0.159 | 0.057 | 0.120 |
| | $(200, 200)$ | 0.055 | 0.091 | 0.045 | 0.069 | 0.039 | 0.059 |
| | $(200, 300)$ | 0.046 | 0.092 | 0.037 | 0.066 | 0.032 | 0.049 |
| | $(300, 300)$ | 0.037 | 0.055 | 0.030 | 0.043 | 0.026 | 0.033 |

Táboa 2: Estimacións por Monte Carlo do $\text{EMCI}_m$ para os modelos LC-1, LC-2, CC-1 e CC-2 con diferentes valores da concentración $\tau$ e tamaños mostrais. Resultados baixo B (benchmark) corresponden aos obtidos con parámetros de suavizado minimizando o $\text{EIC}_m$ e resultados baixo CV refírense aos obtidos con parámetros seleccionados mediante validación cruzada modal.

## 6. ANÁLISE DO ESCAPE DAS LARVAS DE PEIXE CEBRA

Nesta sección analizamos os datos de larvas de peixes cebra presentados na Introdución. Todos os detalles do experimento poden atoparse en Nair et al. (2017b). Os datos represéntanse na Figura 3, tanto no toro coma no plano. O noso obxectivo é estudar como o ángulo no que se aproxima o robot depredador (variable explicativa circular) inflúe á dirección de escape dos peixes (variable resposta circular). Dado que a variable explicativa só cubre unha parte da circunferencia, poderíase argumentar que esta variable pode ser considerada como escalar en lugar de circular. Porén, o soporte da variable explicativa vén dado polo rango de visión de cada animal. Dado que existen animais, como os camaleóns, cun rango de visión de 360 graos, é importante tratar esta variable como circular para que o método sexa extendible a outros animais.

Neste experimento, unha dirección de escape no intervalo $[-\pi, 0)$ indica un escape contralateral, mentres que unha dirección de escape en $[0, \pi)$ clasifícase como ipsilateral. Ademais, valores da dirección de estímulo máis cara a esquerda no panel dereito da Figura 3 indica que o robot se aproximou á larva en cuestión do lado rostral (próximo á parte dianteira do corpo). Por outra banda, direccións de estímulo máis cara a dereita no panel dereito da Figura 3 mostran que o robot se aproximou á larva dende o lado caudal (preto da cola). Aproximarse aos peixes dende os lados rostral ou caudal significa que o robot apareceu na visión periférica dos peixes.

O estimador tipo núcleo da regresión para explicativas e respostas circulares, proposto por Di Marzio et al. (2012), foi aplicado aos datos, onde o parámetro de concentración foi obtido mediante validación cruzada. Dito estimador está representado en verde e con trazo contínuo na Figura 3. De acordo con este estimador, a dirección de escape media é ipsilateral cando o robot se aproxima aos peixes polo lado rostral, contralateral cando se achega ás larvas do seu

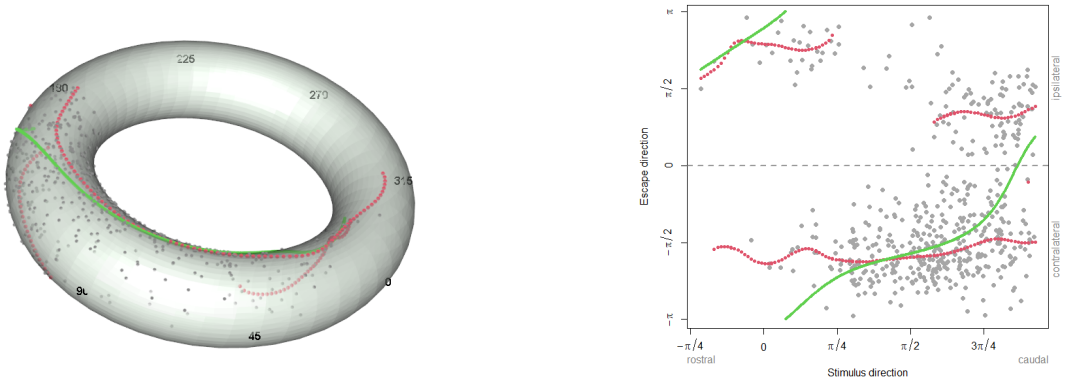



Figura 3: Representacións no toro e no plano dos datos de peixes cebra coa multifunción de regresión estimada (vermello, trazo punteado) e a estimación da función de regresión usual (verde, trazo continuo).

lado esquerdo ou dereito e arredor de cero cando se aproxima polo lado caudal. Co fin de obter máis coñecemento sobre a relación entre a dirección de escape e o ángulo de estímulo, aplicamos o estimador da regresión multimodal, representado en vermello e trazo punteado na Figura 3. Os parámetros de suavizados foron seleccionados mediante o criterio de validación cruzada modal. Podemos observar que, cando o robot aparece na visión periférica dos peixes (tanto o lado rostral como caudal) existen dúas direccións de escape prederidas, unha ipsilateral e outra contralateral. Por outra banda, só se estima unha moda cando o robot aparece dentro do campo de visión frontal dos peixes, indicando un escape contralateral neste caso.

En conclusión, o estimador da regresión multimodal permítenos estimar as dúas direccións de escape prefiridas cando as larvas detectan ao seu depredador mediante a súa visión periférica, é dicir, cando o depredador se achega dende a parte frontal ou traseira. Pola contra, se este se aproxima aos peixes dende os lados dereito ou esquedo dos seus corpos (onde se atopan os ollos dos peixes), a dirección de escape resulta oposta á dirección de ameaza.

## 7. CONCLUSIÓNS

A literatura existente sobre a regresión con variables circulares enfócase predominantemente na regresión *á media*. Con todo, como mostra a nosa análise das direccións de escape de larvas de peixe cebra na Sección 6, cando a densidade condicional dos datos presenta unha estructura multimodal, a estimación da media condicional pode levar a resultados enganosos.

Neste traballo introducimos un método non paramétrico para estudar as modas locais dunha variable resposta condicionadas a unha variable explicativa, cando a resposta e/ou a explicativa son variables circulares. O noso método de regresión multimodal permite estimar os valores máis probables da resposta dada a explicativa, proporcionando un coñecemento máis axeitado da relación entre as variables máis aló da regresión en media. As propiedades do estimador foron estudadas de xeito teórico e tamén mediante un estudo de simulación. O enfoque de suavizado tipo núcleo que se toma neste traballo require da selección de parámetros de suavizado, tarefa que tamén foi realizada.

Dende un punto de vista práctico, na nosa análise dos datos de peixes cebra utilizando a regresión multimoda, vimos que as larvas teñen unha dirección de escape prefirida cando son atacadas dende unha dirección lateral aos seus corpos. Emporiso, cando o depredador se achega aos peixes dende os lados rostral ou caudal, existen dúas direccións de escape prefiridas. Ademais, o estimador multimodal permítenos percibir graficamente as diferenzas entre o comoportamento de escape dos animais cando son perseguidos dende os lados caudal e rostral.

Finalmente, a metodoloxía presentada neste traballo establece as bases para ferramentas inferenciais para regresión con variables circulares, construídas dende unha perspectiva multimodal.

Por exemplo, o estimador introducido na Sección 2 pode utilizarse para construír tanto bandas de confianza como bandas de predicción que, dada a estrutura multimodal dos datos, resultaría en bandas moito máis estreitas que as construídas mediante os estimadores de regresión en media.

## AGRADECEMENTOS

Este traballo foi financiado polo Proxecto MTM2016–76969–P e co-financiado pola European Regional Development Fund (ERDF), os Grupos de Referencia Competitivos 2017–2020 (ED431C 2017/38) da Xunta de Galicia a través da ERDF. O traballo de M. Alonso-Pena foi financiado pola Xunta de Galicia a través do contrato predoutoral con referencia ED481A-2019/139 da Consellería de Educación, Universidade e Formación Profesional. As autoras agradecen tamén ao Centro de Supercomputación de Galicia (CESGA) polos recursos computacionais e a Alejandra López Pérez pola axuda gráfica.

## REFERENCIAS

Ameijeiras-Alonso, J., Lagona, F., Ranalli, M. e Crujeiras, R.M. (2019). A circular nonhomogeneous hidden Markov field for the spatial segmentation of wild fire occurrences. Environmetrics, 30, e2501.

Card, G. e Dickinson, M.H. (2008). Visually mediated motor planning in the escape response of Drosophila. Current Biology, 18, 1300–1307.

Casa, A., Chacón, J.E. e Menardi, G. (2020). Modal clustering asymptotics with applications to bandwidth selection. Electronic Journal of Statistics, 14, 835–856.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. Wiley Interdisciplinary Reviews: Computational Statistics, 10, e1431.

Chen, Y.-C., Genovese, C.R., Tibshirani, R.J. e Wasserman, L. (2016). Nonparametric modal regression. Annals of Statistics, 44, 489–514.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 790–799.

Comaniciu, D. e Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603–619.

Di Marzio, M. and Panzera, A. e Taylor, C.C. (2009). Local polynomial regression for circular predictors. Statistics & Probability Letters, 798, 2066–2075.

Di Marzio, M., Fensore, S., Panzera, A. e Taylor, C.C. (2016). A note on nonparametric estimation of circular conditional densities. Journal of Statistical Computation and Simulation, 86, 2573–2582.

Di Marzio, M., Panzera, A. e Taylor, C.C. (2012). Non-parametric regression for circular responses. Scandinavian Journal of Statistics, 40, 238–255.

Einbeck, J. e Tutz, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speedflow data. Journal of the Royal Statistical Society, Series C, Applied Statistics, 55, 461–475.

Fan, J. e Gijbels, I. (1996). Local Polynomial Modelling and its Applications. Chapman and Hall, London.

Fisher, N.I. (1993). Statistical Analysis of Circular Data. Cambridge University Press, Cambridge.

Fukunaga, K. e Hostetler, L. (1975). Regression models for angular responses. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, 21, 32–40.

Jammalamadaka, S.R. e SenGupta, A. (2001). Topics in Circular Statistics. World Scientific, Singapore.

Kim, S. e SenGupta, A. (2017). Multivariate-multiple circular regression. Journal of Statistical Computation and Simulation, 87, 1277–1291.

Kobayashi, T. e Otsu, N. (2010). Von mises-fisher mean shift for clustering on a hypersphere. In Proceedings of the 20th international conference on pattern recognition, pages 2130–2133. IEEE.

Ley, C. e Verdebout, T. (2017). Modern Directional Statistics. Chapman & Hall, Boca Raton.

Marchetti, M. e Scapini, F. (2003). Use of multiple regression models in the study of sandhopper orientation under natural conditions. Estuarine, Coastal and Shelf Science, 58, 207–215.

Mardia, K.V. e Jupp, P.E. (2000). Directional Statistics. John Wiley & Sons, Inc., New York.

Mooney, J.A., Helms, P.J. e Jollife, I.T. (2003). Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. Computational Statistics and Data Analysis, 41, 505–513.

Nair, A., Changsing, K., Stewart, W.J. e McHenry, M.J. (2017a). Data from: Fish prey change strategy with the direction of a threat. https://doi.org/10.5061/dryad.47mq9.

Nair, A., Changsing, K., Stewart, W.J. e McHenry, M.J. (2017b). Fish prey change strategy with the direction of a threat. Proceedings of the Royal Society B, 284, 20170393.

Oba, S., Kato, K. e Ishii, S. (2005). Multi-scale clustering for gene expression profiling data. In 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), pages 210–217. IEEE.

Obleser, P., Hart, V. Malkemper, E.P. et al. (2016). Compass-controlled escape behavior in roe deer. Behavioral Ecology and Sociobiology, 70, 1345–1355.

Oliveira, M., Crujeiras, R.M. e Rodríguez-Casal, A. (2013). Nonparametric circular methods for exploring environmental data. Environmental and Ecological Statistics, 20, 1–17.

Pewsey, A., Neuhuser, M. e Ruxton, G.D. (2013). Directional Statistics. Oxford University Press, Oxford.

Sato, N., Shidara, H. e Ogawa, H. (2019). Trade-off between motor performance and behavioural flexibility in the action selection of cricket escape behaviour. Scientific Reports, 9, 18112.

Scapini, F., Aloia, A., Bouslama, M. et al. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed mediterranean beach. Behavioral Ecology, 51, 403–414.

Scott, D.W. (1992). Multivariate Density Estimation. Wiley, New York

SenGupta, S. e Rao, J.S. (1966). Statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Godavari Valley. Sankhya: The Indian Journal of Statistics, Series B, 28, 165–174.

Wikimedia Commons (2008). Larval zebrafish. Author: CSIRO. https://commons.wikimedia.org/wiki/File:CSIRO_ScienceImage_7598_larval_zebra.jpg [Online; accessed July 27th, 2021].

Wikimedia Commons (2014). Life cycle of Zebrafish. https://commons.wikimedia.org/wiki/File:Zebrafish-model-4-638.jpg [Online; accessed July 27th, 2021].

Zhang, Y. e Chen, Y.-C. (2020). Kernel smoothing, mean shift, and their learning theory with directional data. arXiv e-prints, page arXiv:2010.13523.

Zhou, H. e Huang, X. (2016). Nonparametric modal regression in the presence of measurement error. Electronic Journal of Statistics, 10, 3579–3620.

Zhou, H. e Huang, X. (2019). Bandwidth selection for nonparametric modal regression. Communications in Statistics - Simulation and Computation, 48, 968–984.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# A DISTANCE COVARIANCE APPROACH TO GENOME-WIDE ASSOCIATION STUDIES

Fernando Castro-Prado[1], Dominic Edelmann[2] and Jelle J. Goeman[3]

[1]Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela
[2]German Cancer Research Center (DKFZ), Heidelberg, Germany.
[3]Leiden University Medical Center (LUMC), Leiden, the Netherlands.

## ABSTRACT

Testing the nonparametric hypothesis of independence consists in assessing whether a group of two or more variables of interest are associated to one another or not. A relevant problem in biomedicine that boils down to that statistical question are case-control studies of complex human traits (GWASs). These involve the usage of data from single nucleotide polymorphisms (SNPs) as discrete covariates in a high-dimensional and low-sample-size setting; as well as a univariate response (e.g., absolutely continuous) of phenotypical nature, whose association —or lack thereof— with the different SNPs is to be tested for.

Generalized distance correlation is an association measure that characterizes general statistical independence between random variables, which can be extended to metric, semimetric and even premetric spaces. By applying this technique to the GWAS problem, one can obtain a flexible interpretation of the values of the support space for each SNP, treating them as arbitrary objects. This theory is mathematically equivalent to the Hilbert–Schmidt independence criterion that is popular in the machine learning community and, therefore, which in turn allows all of it to be rewritten in the language of linear global tests.

In the current work we present generalized distance covariance as a novel tool for approaching GWASs, which have been of maximum interest in genomics for the last decade. The main theoretical results on the present document have to do with allowing for the discoveries of general models of interaction —not only additive effects—, with the option to select a priori the desired model. In addition, we derive the asymptotic distribution of the test statistic in explicit terms, thus speeding up the usually time-consuming distance correlation procedure (i.e., avoiding resampling schemes). Finally we present some further practical considerations and data analysis examples.

**Keywords**: Association measures; Complex disease genomics; Distance correlation; Global tests; Hilbert–Schmidt independence criterion; Statistical learning.

## 1. INTRODUCTION

The genetic variability of biological quantitative traits responds to the combined effect of a large number of variants along the genome. To identify them, genome-wide association studies (GWASs) tend to focus on the ones that are common in the population (due to higher power and relative ease of imputation), which altogether have turned out to polygenically explain a considerable portion of the overall trait heritability (Visscher et al., 2009).

GWASs involve testing genetic variants across the genomes of many sampled (human) individuals to identify genotype-phenotype associations. Thus, the response variable corresponds to a phenotypic characteristic of interest, which can be binary (typically presence/absence of a complex disease) or continuous (e.g., physical measures of the human body, concentration of certain

molecules in the blood, cardiological parameters, indicators of severity or stage of an illness, age at which a body development hallmark is achieved, and so forth). Henceforward we will restrict ourselves to the latter case. Whereas the binary scenario requires two groups (called "cases" and "controls"), for our interests, a single large pool of individuals will be enough (Zhang et al., 2018).

A GWAS database typically offers observations of single-nucleotide polymorphisms (SNPs) with a quantitative response. For this purpose, there is a need for computationally efficient powerful testing methodology, taking into account the particular structure of the data. In most applications, a standard linear model is applied, regarding the state values $\{0, 1, 2\}$ at a SNP as either categorical or continuous. However it has been shown that these approaches often lead to suboptimal results, while it seems to be preferable to combine different models of genotype-phenotype association (Lettre et al., 2007). The methodology for computing critical values for these combinations is in turn substantially more expensive, from a computational point of view.

In this paper, we present a novel method for testing the association of a single SNP with a quantitative response based on generalized distance covariance (Sejdinovic et al., 2013; Edelmann and Goeman, 2021). This method is computationally very efficient and has a model-based interpretation. Moreover, it features an optimality property, namely that the test is locally most powerful under the corresponding model assumptions.

## 2. BACKGROUND

### 2.1. Models for the association between SNPs and quantitative traits

In the present work, we study methods for analyzing associations between genotypes and quantitative traits. A certain position in the genome of an organism is called a *locus*, the piece of DNA at that locus is called an *allele*. The genomes of organisms of the same species mostly coincide. When studying the association of genotypes with quantitative, one hence considers the loci where two different alleles are present in the population.

We will consider that the loci of interest are positions of single nucleotides on the genome, where two different variants (base-pairs) are present in the population. These variants are called *single-nucleotide polymorphisms*. Our goal is to develop testing methodology for assessing the association of a single SNP with a quantitative trait.

Humans are diploid organisms, i.e., they have two matching sets of chromosomes and either of the two alleles may be present in each of the two sets. Hence, we distinguish between three different states: $A_1A_1$ (the first allele is present in both sets), $A_1A_2$ (the first allele is present in one set and the second allele in the other set) and $A_2A_2$ (the second allele is present in both datasets). The state of the genotype will be modelled by a random element $X$ with support $\{0, 1, 2\}$, which encodes the count of the second allele. It is common convention to define $A_2$ as the minor allele, i.e. the allele that is less frequent in the population; nonetheless the methodology presented in this paper will not rely on this convention.

| | $X = 0\,(A_1A_1)$ | $X = 1\,(A_1A_2)$ | $X = 2\,(A_2A_2)$ |
|---|---|---|---|
| mean | $\mu_0$ | $\mu_1$ | $\mu_2$ |
| standardized effect (for $\mu_0 \neq \mu_2$) | 0 | $h$ | 1 |

Table 1: Association models between SNPs and quantitative traits.

For studying different models between the state $X \in \{0, 1, 2\}$ of a certain SNP and an absolutely continuous response $Y \in \mathbb{R}$, let us define $\mu_j = \mathbb{E}[Y|X = j]$, for $j \in \{0, 1, 2\}$. The associations between $X$ and $Y$ that are classically considered (Gillespie, 2004) assume that the means of the two homozygous states are different $\mu_2 \neq \mu_0$, whereas the mean of the heterozygous state may either coincide with one the homozygous states or be different. In this setting, the relations between $X$ and $Y$ are summarized on Table 1. The standardized effect for each state $j \in \{0, 1, 2\}$ is hereby calculated as $\frac{\mu_j - \mu_0}{\mu_2 - \mu_0}$. The association models are then classified based on the biological interpretation of the value of $h := \frac{\mu_1 - \mu_0}{\mu_2 - \mu_0} \in \mathbb{R}$:

- $h = 0$: *dominant-recessive* model; where $A_1$ is *dominant*, $A_2$ is *recessive*.

- $h = 1$: dominant-recessive model; where $A_2$ is dominant, $A_1$ is recessive.
- $h \in (0, 1)$: *codominant model*; a codominant model with $h = \frac{1}{2}$ is called *additive* model.
- $h < 0$ or $h > 1$: *overdominant* model.

The parameter $h$ is called the *heterozygous effect*. In the course of this paper, we will also consider models for which $\mu_0 = \mu_2$ and $\mu_1 = \mu_0$, we will refer to such a model as *purely heterozygous model*.

### 2.2. Generalized distance covariance

Distance covariance and distance correlation are novel dependence measure that have been derived by Székely et al. (2007); Székely and Rizzo (2009) as alternatives to the classical covariance and Pearson correlation. Given random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with finite second moments, distance covariance can be defined as

$$\begin{aligned}\mathcal{V}^2(X,Y) = \mathbb{E}\Big[\|X - X'\|\|Y - Y'\|\Big] + \mathbb{E}\Big[\|X - X'\|\Big]\,\mathbb{E}\Big[\|Y - Y'\|\Big] \\ - 2\mathbb{E}\Big[\big(\|X - X'\|\big)\big(\|Y - Y''\|\big)\Big],\end{aligned} \tag{2.1}$$

where $(X', Y')$ and $(X'', Y'')$ denote i.i.d. copies of $(X, Y)$.

In this paper we will consider a version of *generalized distance covariance*, which has been derived in Sejdinovic et al. (2013); Edelmann and Goeman (2021).

For this purpose, we will call a function $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ on a set $\mathcal{Z}$ a premetric if it is symmetric in its arguments and satisfies $\rho(z, z) = 0$ for all $z \in \mathcal{Z}$. Then $(\mathcal{Z}, \rho)$ is called a premetric space.

A premetric space $(\mathcal{Z}, \rho)$ is said to have negative type (Sejdinovic et al., 2013, Definition 2), if for all $n \geq 2$, $z_1, \ldots, z_n \in \mathcal{Z}$ and $a_1, \ldots, a_n \in \mathbb{R}$ with $\sum_{i=1}^n a_i = 0$,

$$\sum_{i,j=1}^{n} a_i a_j \rho(z_i, z_j) \leq 0.$$

Now let $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ denote premetrics of negative type on $\mathcal{X}$ and $\mathcal{Y}$. Then, the *generalized distance covariance* is defined as follows.

$$\begin{aligned}\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = \mathbb{E}(\rho_{\mathcal{X}}(X,X')(\rho_{\mathcal{Y}}(Y,Y') - \rho_{\mathcal{Y}}(Y,Y'') \\ - \rho_{\mathcal{Y}}(Y',Y'') + \rho_{\mathcal{Y}}(Y'',Y''')),\end{aligned} \tag{2.2}$$

where $(X, Y), (X', Y'), (X'', Y''), (X''', Y''')$ are i.i.d. copies of $(X, Y)$.

Consider now i.i.d. joint samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ of $X$ and $Y$, and define the $(i, j)$th element of the distance matrices $\boldsymbol{D}^{\boldsymbol{X}}$ and $\boldsymbol{D}^{\boldsymbol{Y}}$ by $\boldsymbol{D}^{\boldsymbol{X}}_{i,j} = \rho_{\mathcal{X}}(X_i, X_j)$ and $\boldsymbol{D}^{\boldsymbol{Y}}_{i,j} = \rho_{\mathcal{Y}}(Y_i, Y_j)$, respectively. Then defining the double centered versions

$$\tilde{\boldsymbol{D}}^{\boldsymbol{X}} = (I - H)\boldsymbol{D}^{\boldsymbol{X}}(I - H), \qquad \tilde{\boldsymbol{D}}^{\boldsymbol{Y}} = (I - H)\boldsymbol{D}^{\boldsymbol{Y}}(I - H)$$

where $H = \frac{1}{n}\mathbf{1}\mathbf{1}^t$, a consistent empirical estimator for (2.2) is given by Székely et al. (2007),

$$\widehat{\mathcal{V}}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} \tilde{\boldsymbol{D}}^{\boldsymbol{X}}_{i,j}\tilde{\boldsymbol{D}}^{\boldsymbol{Y}}_{i,j} = \frac{1}{n^2}\mathrm{tr}\,(\boldsymbol{D}^{\boldsymbol{X}}\tilde{\boldsymbol{D}}^{\boldsymbol{Y}}) = \frac{1}{n^2}\mathrm{tr}\,(\tilde{\boldsymbol{D}}^{\boldsymbol{X}}\boldsymbol{D}^{\boldsymbol{Y}}). \tag{2.3}$$

A particular interesting choice for $\rho_{\mathcal{Y}}$ for real-valued $Y$ ($\mathcal{Y} = \mathbb{R}$) is one-half of the squared Euclidean distance $\rho_{\mathcal{Y}}(y, y') = \frac{1}{2}|y - y'|^2$, in which case $\widehat{\mathcal{V}}^2(\boldsymbol{X}, \boldsymbol{Y})$ can be interpreted as the locally most powerful test statistic in certain Gaussian regression models (Edelmann and Goeman, 2021, Theorem 3).

## 3. GENERALIZED DISTANCE COVARIANCE FOR TESTING THE ASSOCIATION OF A SINGLE SNP WITH A CONTINUOUS RESPONSE

### 3.1. Tailoring premetrics to GWAS data

Our goal will be to define versions of generalized distance covariance $\mathcal{V}_{\rho_\mathcal{X},\rho_\mathcal{Y}}$ for testing independence between a SNP $X \in \mathcal{X} := \{0,1,2\}$ and a quantitative response $Y \in \mathcal{Y}$. In order to obtain interpretations of the test statistics as locally most powerful tests and to facilitate generalizations, we will always apply $\rho_\mathcal{Y}(y,y') = \frac{1}{2}|y-y'|^2$.

For defining meaningful distances on the support space of the SNPs, we note that 0 and 2 correspond to homozygous states, while 1 denotes the heterozygous state. Since the notation for the two homozygous states is arbitrary and may be interchanged, it is reasonable to only consider symmetric distances, which yields $d(0,1) = d(1,2) = 1$.

The resulting family of distances is characterized by the positive real number $b := d(0,2)$. Moreover, for a semimetric to define a distance covariance in the sense of Sejdinovic et al. (2013), it must be of negative type and for this in turn, its square root must be a metric (p. 2266 of the aforementioned article). Hence $b$ must satisfy $b \leq 2^2 = 4$. Proposition 3 in Sejdinovic et al. (2013) implies that $b \in (0,4]$ defines indeed valid semimetrics of negative type. Without loss of generality, one can assimilate points that are separated with distance zero and therefore extend the theory to premetrics, allowing for the case $b = 0$ (i.e., dropping the identity of indiscernibles).

We will hence study the family of premetrics $\{d_b\}_{b\in[0,4]}$, where $d_b : \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ is such that $d_b(0,1) = d_b(1,2) = 1$ and $d_b(0,2) = b \in [0,4]$. Important special cases are:

1. The discrete metric $d_1(0,1) = d_1(1,2) = d_1(0,2) = 1$, as studied by Castro-Prado et al. (2020+).
2. The Euclidean metric $d_2(x,x') = |x-x'|$, connected to standard distance covariance on the ordered set $\{0,1,2\} \subset \mathbb{R}$.
3. The squared Euclidean distance $d_4(x,x') = (x-x')^2$, linked to linear regression on the ordered set $\{0,1,2\} \subset \mathbb{R}$.
4. Similarly, any premetric $d_b$ with $b \in (1,4)$ is related to the $\alpha-$distance covariance of (Székely et al., 2007) for $\alpha = \log_2 b$.

For the rest of the article, we use the simplified notation

$$\mathcal{V}_b := \mathcal{V}_{d_b,\rho_\mathcal{Y}}, \quad \widehat{\mathcal{V}}_b := \widehat{\mathcal{V}}_{d_b,\rho_\mathcal{Y}},$$

where $\rho_\mathcal{Y}$ is one half of the squared Euclidean distances. Unlike classical distance covariance, $\mathcal{V}_b$ does not characterize independence because $\rho_\mathcal{Y}$ is not of strong negative type. However, $\mathcal{V}_b$ can detect all associations defined via the classical phenotype-genotype association models introduced in Subsection 2.1. For this purpose, we again consider

$$\mu_j = \mathbb{E}[Y|X=j]$$

for $j \in \mathcal{X} \equiv \{0,1,2\}$. Then, under some regularity conditions, we have that the distance covariance between $X$ and $Y$ vanishes if and only if the mean effects of $Y$ are homogeneous among the categories of $X$: $\mu_0 = \mu_1 = \mu_2$.

**Theorem 3.1.** *Let $\mathbb{E}[Y^2] < \infty$ and. If $\mu_0 = \mu_1 = \mu_2$, then*

$$\mathcal{V}_b^2(X,Y) = 0.$$

*Moreover, if $b \in (0,4)$ and $p_j > 0$ for $j \in \{0,1,2\}$, then $\mu_i \neq \mu_j$ for some $i \neq j$ implies that*

$$\mathcal{V}_b^2(X,Y) > 0.$$

**Remark 3.2.** *For the limit cases $b \in \{0,4\}$, we can always find a configuration with $\mu_i \neq \mu_j$ for some $i \neq j$, but $\mathcal{V}_b^2(X,Y) = 0$.*

In the following, we establish tests for the null hypothesis

$$H_0 : \mu_0 = \mu_1 = \mu_2$$

using the estimator $\widehat{V}_b^2$.

### 3.2. Testing

The problem of efficient $p$-value calculation has been one of the main sources of criticism of classical distance covariance. Notably, in the original work by Székely et al. (2007) a permutation approach has been proposed, which is virtually unfeasible for testing a large number of SNPs. In the setting of the generalized distance covariance $\mathcal{V}_b^2$, we can derive a closed form expression for the limit of the test statistic, which can be efficiently calculated with standard statistical software (Hu, 2020).

**Theorem 3.3.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ denote i.i.d. samples of jointly distributed random variables $(X, Y) \in \{0, 1, 2\} \times \mathbb{R}$ with $\mathbb{E}[Y^2] = \sigma_Y^2 < \infty$. If $X$ and $Y$ are independent, then, for $n \to \infty$,*

$$n\,\widehat{\mathcal{V}}_b^2 \xrightarrow{\mathcal{D}} \sigma_Y^2(\lambda_1 Z_1^2 + \lambda_2 Z_2^2),$$

*where $Z_1^2$ and $Z_2^2$ are chisquare distributed with one degree of freedom and $\lambda_1$ and $\lambda_2$ are the eigenvalues of the matrix*

$$K = \begin{pmatrix} \frac{b}{2}(p_0 + p_2 - (p_0 - p_2)^2) & \sqrt{\frac{b\,(4-b)}{4}}(-p_1(p_0 - p_2)) \\ \sqrt{\frac{b\,(4-b)}{4}}(-p_1(p_0 - p_2)) & \frac{4-b}{2}(p_1 - p_1^2) \end{pmatrix},$$

## 4. LOCALLY MOST POWERFUL PROPERTY AND INTERPRETATIONS

The classical score test (Cox and Hinkley, 1979) for a model with likelihood $\ell^*(\theta; \boldsymbol{Z})$ where $\boldsymbol{Z} \in \mathbb{R}^n$ is an observation and $\theta \in \Theta \subset \mathbb{R}$ is a univariate parameter, is a one-sided test $H_0^* : \theta = \theta_0$ against $H_1^* : \theta > \theta_0$ that rejects $H_0^*$ if

$$S^* = \frac{d \log \ell^*(\theta_0; \boldsymbol{Z})}{d\theta} \geq c$$

for some critical value $c$. The score test is also known as the *locally most powerful test* since it satisfies the following optimality property

**Lemma 4.1** (Goeman et al. (2006), Lemma 2)**.** *For $\theta \in \Theta$, denote by $Z_\theta \in \mathbb{R}^n$ a random variable distributed corresponding to $\ell^*(\theta; \boldsymbol{Z})$ and denote its probability measure by $P_{Z_\theta}$. Suppose that the derivative $\frac{d\ell^*(\theta;\boldsymbol{Z})}{d\theta}$ exists for all $Z \in \mathbb{R}^n$ and is bounded in a neighbourhood of $\theta_0$. Then, for any test $H_0^*$ with critical region $A$ and power function $w(\theta) = P_{Z_\theta}(A)$, the derivative $\frac{dw(\theta_0)}{d\theta}$ exists. Also, denote the power function of the score test statistic by $w^*(\theta) = P_{Z_\theta}(S^* \geq c)$ for some $c \geq 0$. Then*

$$w(\theta_0) \leq w^*(\theta_0)$$

*implies*

$$\frac{d}{d\theta} w(\theta_0) \leq \frac{d}{d\theta} w^*(\theta_0).$$

Since for small $h$

$$P_{\theta_0 + h}(A) \approx P_\theta(A) + h\, w(\theta_0),$$

this implies that the score test is the most powerful test for detecting local alternatives corresponding to infinitesimally small deviations from $\theta_0$.

Edelmann and Goeman (2021) have shown that, if the squared Euclidean distance is applied on the response, the generalized distance covariance arises from the score test statistic in certain Gaussian regression models. This implies that $\widehat{V}_b^2$ has an interpretation as locally most powerful test statistic, which we state in Theorem 4.2 and Remark 4.3.

**Theorem 4.2.** *Let $(\phi_1, \ldots, \phi_r)$ be a feature map of the distance $d_b$. Consider the model*

$$y_i = \sum_{j=1}^{r} \beta_j \phi_j(x_i) + \mu + \varepsilon,$$

*where $\mu$ is known, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\beta_j = \tau B$ with $\tau \in \mathbb{R}$ and $B$ is a random variable with $E[B] = 0$ and $E[BB^t] = I_r$. Then the locally most powerful test statistic for testing*

$$H_0 : \tau^2 = 0 \text{ against } H_1 : \tau^2 > 0$$

*is given by*

$$\widehat{T}_b = -\frac{1}{n^2} \sum_{i,j=1}^{n} d_b(x_i, x_j)(y_i - \mu)(y_j - \mu). \tag{4.1}$$

**Remark 4.3.** *The population mean $\mu$ is typically unknown in practice. By inserting the sample mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ for $\mu$ in (4.1), we see that a pivot statistic for $\widehat{T}_b$ is given by the generalized distance covariance $\widehat{\mathcal{V}}_b$.*

We note that in GWAS, it is usually conjectured that the effect of a single SNPs on a quantitative trait is small, hence the assumption of a small $\tau$ appears sensible. Consequently the locally most powerful property is particularly desirable for this setting.

### 4.1. Interpretation as locally most powerful test for situations, where dominant-recessive, additive or purely heterozygous models are expected for certain fractions of SNPs

The marginal distribution of $B$ is not specified in Theorem 4.1. There are several choices of $B$ that lead to interesting model-based interpretations. A first interpretation is provided by Corollary 4.4.

**Corollary 4.4.** *Let $(\phi_1, \ldots, \phi_r)$ be a feature map of the distance $d_b$ and, for $j \in \{1, \ldots, r\}$, let $c_j > 0$ be arbitrary constants and denote $\psi_j(\cdot) = \phi_j(\cdot)/c_j$. Let $U$ be a discrete random variable with $P(U = j) = \frac{c_j^2}{\sum_{i=1}^n c_i^2}$ and consider the model*

$$y_i = \tau\, A \sum_{j=1}^{r} 1_{\{U=j\}} \psi_j(x_i) + \mu + \varepsilon,$$

*where $\tau \in \mathbb{R}$, $\mu$ is known, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $A$ is a random variable with $\mathbb{E}[A] = 0$ and $\mathbb{E}[A^2] = 1$ (e.g. $P(A = 1) = P(A = -1) = \frac{1}{2}$). Then the locally most powerful test is given by (4.1).*

Corollary 4.4 states that $\mathcal{V}_b$ can be regarded as a pivot to the locally most powerful test statistic in a model, where each of $r$ different association patterns between $X$ and $Y$ (specified by the $r$ features of the feature maps) is present with a certain probability. We note that this is different from a mixture model, in the sense that the hyperparameter $U$ does not depend on $i$ and hence the same model is true for all samples $i$.

Considering that we would usually perform the test based on $\mathcal{V}_b$ repeatedly to test a large number of SNPs, this implies that the corresponding test is optimal for situations in which each of the $r$ different association patterns expressed by $\psi_j$ shows up for a fraction of $\frac{c_j^2}{\sum_{i=1}^n c_i^2}$ SNPs. Applying this corollary with different feature maps yields different interesting interpretations of the models where $\mathcal{V}_b$ performs well.

As we have pointed out in Section 3, every distance $d_b$ can be expressed as a feature map with two features. For interpretation reasons, it is however useful to consider feature maps with more than two features. For $b \in [0, 2]$, an interesting feature map is:

$$\phi_1 = \sqrt{b} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \phi_2 = \sqrt{b} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \phi_3 = \sqrt{2-b} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Applying Corollary 4.4 now yields, that, for $b \in [0, 2]$, $\mathcal{V}_b$ can be interpreted as pivot to the locally most powerful test for situations where we expect a dominant-recessive model with probability $\frac{2\,b}{2+b}$ (for which each of the homozygous states is dominant with equal probability) and a purely heterozygous model with probability $\frac{2-b}{2+b}$. In particular, $\mathcal{V}_2$ corresponds to the locally most powerful test in a random dominant-recessive model, for which each of the homozygous states is dominant

with probability $\frac{1}{2}$. On the other hand, $\mathcal{V}_1$ corresponds to the situation in which, for each $j \in \{0,1,2\}$, $\mu_j$ differs from the other two means with probability $\frac{1}{3}$. For $\mathcal{V}_0$, $U$ is constant; so this yields the locally most powerful test for a purely heterozygous model.

Similarly, for $b \in [2,4]$, we can derive the feature map

$$\phi_1 = \sqrt{4-b}\begin{pmatrix}0\\0\\1\end{pmatrix}, \quad \phi_2 = \sqrt{4-b}\begin{pmatrix}0\\1\\1\end{pmatrix}, \quad \phi_3 = 2\sqrt{b-2}\begin{pmatrix}0\\\frac{1}{2}\\1\end{pmatrix}$$

Applying Corollary 4.4 now yields that for $b \in [2,4]$, $\mathcal{V}_b$ is the locally most powerful test under the assumption that the absolute difference between the two homozygous states is $|\tau|$ and the heterozygous state takes the value of each of the homozygous states with probability $\frac{4-b}{2b}$ and the average of the two values with probability $\frac{4(b-2)}{2b}$. This corresponds to the situation, where we expect dominant-recessive and additive models, each with a certain probability. The extreme case $\mathcal{V}_4$ corresponds to the locally most powerful test in a purely additive model.

### 4.2. Interpretation as the locally most powerful test for codominant and overdominant models with random parameters

While it is common that the response values for the heterozygous state lie between the values of the two homozygous states, it is rather unlikely that we encounter an exact additive model. Instead, the response values of the heterozygous state will often lie closer to one of the homozygous states. A model which assumes that the response values for the heterozygous state lie somewhere between the response values of the two homozygous states is referred to as a *codominant model*, as previously indicated.

We will now show that, for $b \in (2,4]$, $\mathcal{V}_b^2$ can be interpreted as the locally most powerful test statistic in certain random codominant models. For $b \in [0,2)$, we obtain a similar interpretation based on overdominant models. For this purpose, we first state the following alternative formulation of the locally most powerful property.

**Theorem 4.5.** *Consider the distance $d_b$ and assume the model*

$$y_i = \begin{cases}\mu+\varepsilon, & \text{if } x_i = 0,\\ \mu+\beta_1+\varepsilon, & \text{if } x_i = 1\\ \mu+\beta_1+\beta_2+\varepsilon & \text{if } x_i = 2,\end{cases}$$

*where $\mu$ is known, $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ and $\beta_j = \tau B$ with $\tau \in \mathbb{R}$ and $B$ is a random variable with $E[B]=0$ and*

$$E[BB^t] = c\begin{pmatrix}1 & \frac{b}{2}-1\\ \frac{b}{2}-1 & 1\end{pmatrix},$$

*where $c$ is some constant. Then the locally most powerful is given by* (4.1).

This gives the interesting interpretation of $\widehat{\mathcal{V}}_b$ as a pivot of the locally most powerful test statistic in a regression models with correlated regression parameters. For $b \in [0,2)$, the correlation between $\beta_1$ and $\beta_2$ is negative. In this case, we can choose $B$ in a way such that $\beta_1$ and $\beta_2$ always have the opposite sign. For $b \in (2,4]$ on the other hand, the correlation between $\beta_1$ and $\beta_2$ is positive and hence we can choose $B$ in a way such that $\beta_1$ and $\beta_2$ always have the same sign.

Reminding us of the different association models introduced in Section 2.1, we can interpret $\mathcal{V}_b$ with $b \in (0,2)$ as the locally most powerful test in an overdominant model with random heterozygous effect. Analogously $\mathcal{V}_b$ with $b \in (2,4)$ can be interpreted as the locally most powerful test in an codominant model with random heterozygous effect.

For constructing bivariate random vectors with zero mean for which the marginals have equal or opposite sign, we note that for any pair of non-negative random variables $A = (A_1, A_2)^t$ with, we can use a random variable $U$ with $P(U=1) = P(U=-1) = \frac{1}{2}$ to construct mean zero random variables

$$B = \begin{pmatrix}B_1\\B_2\end{pmatrix} = \begin{pmatrix}UA_1\\UA_2\end{pmatrix} \quad \widetilde{B} = \begin{pmatrix}\widetilde{B}_1\\\widetilde{B}_2\end{pmatrix} = \begin{pmatrix}UA_1\\-UA_2\end{pmatrix}$$

Then the marginals of $B$ are of equal sign, the ones of $\widetilde{B}$ have opposite sign and

$$\mathrm{Cor}(B_1, B_2) = \frac{\mathbb{E}[A_1 A_2]}{\sqrt{\mathbb{E}[A_1^2]\mathbb{E}[A_2^2]}} \qquad \mathrm{Cor}(B_1, B_2) = -\frac{-\mathbb{E}[A_1 A_2]}{\sqrt{\mathbb{E}[A_1^2]\mathbb{E}[A_2^2]}}$$

A particularly interesting choice for interpreting $\mathcal{V}_b$ with values $b \in (2, 4)$ is presented in the following corollary.

**Corollary 4.6.** *Consider distance $d_b$ and assume the model*

$$y_i = \begin{cases} \mu + \varepsilon, & \text{if } x_i = 0, \\ \mu + \tau H + \varepsilon, & \text{if } x_i = 1 \\ \mu + \tau + \varepsilon & \text{if } x_i = 2, \end{cases}$$

*where $\mu$ is known, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and the heterozygous effect $H$ is beta-distributed with parameters $(\frac{b-2}{4-b}, \frac{b-2}{4-b})$ Then the locally most powerful is given by* (4.1).

Corollary 4.6 states that, for $b \in (2, 4)$, $\mathcal{V}_b$ is the optimal test in a codominant model for which the heterozygous effect parameter $H$ is beta-distributed. A notable special case is $b = 3$, for which $H$ is uniformly distibuted.

## 5. APPLICATION AND PRACTICAL CONSIDERATIONS

### 5.1. Meaning and choice of $b$

Table 2: Genetic model against which $\mathcal{V}_b$ provides the locally most powerful test, for different values of $b \in [0, 4]$.

| | Genetic modell |
|---|---|
| $b = 0$ | purely heterozygous model |
| $b \in (0, 1)$ | overdominant model with large (random) heterozygous effect $h$ |
| $b = 1$ | agnostic model |
| $b \in (1, 2)$ | overdominant model with small (random) heterozygous effect $h$ |
| $b = 2$ | dominant-recessive model |
| $b \in (2, 3)$ | codominant model where $h$ tends to be close to 0 or 1 ($h \sim \beta(\frac{b-2}{4-b}, \frac{b-2}{4-b})$) |
| $b = 3$ | codominant model, heterozygous effect $h$ is uniformly distributed on $0, 1$ |
| $b \in (3, 4)$ | codominant model where $h$ tends to be close to $\frac{1}{2}$ ($h \sim \beta(\frac{b-2}{4-b}, \frac{b-2}{4-b})$) |
| $b = 4$ | additive model |

We have obtained model-based interpretations of the test statistic that are summarized in Table 5.1 We emphasize again that all choices of $b \in (0, 4)$ are consistent against all alternatives; only the degenerate cases $b \in \{0, 4\}$ do not guarantee that. Most known associations in genetics follow either a dominant-recessive or a codominant model. Hence, it seems reasonable for most applications to assume $b \in [2, 4]$.

Interpreting $\mathcal{V}_b$ as a mixture of additive and dominant recessive models, we easily calculate that $b = \frac{12}{5} = 2.4$ gives the test statistic for the setting, where we assume a dominant, recessive and additive model with probability $\frac{1}{3}$ each. On top of that, $b = \frac{8}{3} \approx 2.67$ relates to the situation of a dominant and recessive model with probability $\frac{1}{4}$ each and an additive model with probability $\frac{1}{2}$. Finally, $b = 3$ is optimal for the setting where the heterozygous effect is uniformly distributed on the interval $[0, 1]$.

Considering that both codominant models and dominant-recessive models frequently arise in practice, it appears that $b \in [2, 3]$ is a good choice for most applications. The impact of parameter $b$ on the power of the test is further investigated using simulations in Section 5.2.

### 5.2. Simulation study

We investigate the effect of parameter $b$ for different association models between a SNP $X$ and a quantitative response $Y$. For simplicity, the minor allele frequency for this SNP is fixed at 0.5.

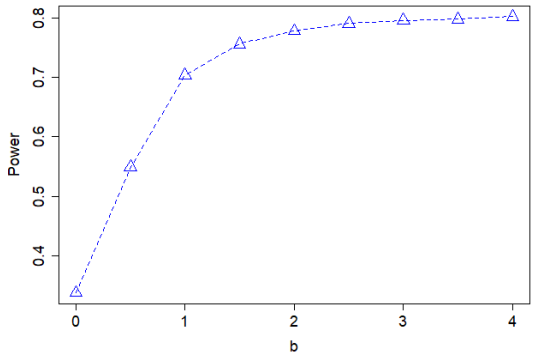



Figure 1: Power of the generalized distance covariance test based on $\mathcal{V}_b$ for the additive model in (5.1)

.

The regression coefficients $\beta$ are chosen in a way that the power for the most powerful test in every model is approximately 0.8, the error term $\varepsilon_i$ follows a standard normal distribution, the sample size is $n = 100$. All tests are performed with a significance level of $\alpha = 0.05$

We consider the following models

1. An additive model,
$$y_i = \beta X_i + \varepsilon, \tag{5.1}$$
with $\beta = 0.5$ (Figure 1).

2. A codominant model,
$$h\beta\, 1_{\{X=1\}} + \beta\, 1_{\{X=2\}} + \varepsilon \tag{5.2}$$
with $\beta = 0.8$, $h = 0.25$. (Figure 2).

3. A dominant-recessive model,
$$y_i = \beta\, 1_{\{X=2\}} + \varepsilon, \tag{5.3}$$
with $\beta = 0.75$ (Figure 3).

4. An overdominant model,
$$y_i = h\beta\, 1_{\{X=1\}} + \beta\, 1_{\{X=2\}} + \varepsilon, \tag{5.4}$$
with $\beta = 0.5$, $h = 1.5$. (Figure 4)

5. A purely heterozygous model,
$$y_i = \beta\, 1_{\{X=2\}} + \varepsilon, \tag{5.5}$$
with $\beta = 0.58$ (Figure 5)

If the true model is dominant-recessive or codominant, no choice in $b \in [1, 4]$ shows substantial loss of power against the corresponding best method. A bit surprising may be the flat shape of the power curve for the codominant model showing no particular advantage for any choice in $[2, 4]$. Overall, $b = 2$ seems to be a quite robust choice, with only minimal power losses (compared to the best $b$) for the codominant and additive model.

For overdominant models, the performance of the large values of $b$ plummets, even in cases where the heterzygous effect $h$ is only slightly larger than 1 (Figure 4), whereas $b \in \{1.5, 2\}$ still shows satisfactory performance. For the case of a purely heterozygous model, the power of $\mathcal{V}_4$ coincides with the nominal level of 0.05. Clearly, $b = 2$ shows better performance than higher values of $b$ in these scenarios.

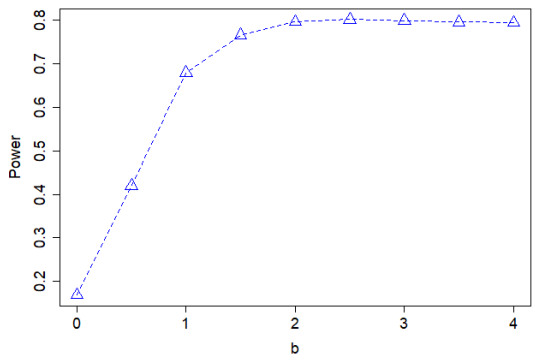



Figure 2: Power of the generalized distance covariance test based on $\mathcal{V}_b$ for the codominant model in (5.2)

.

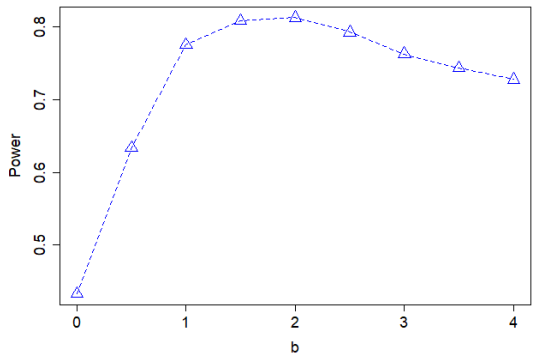



Figure 3: Power of the generalized distance covariance test based on $\mathcal{V}_b$ for the dominant-recessive model in (5.3)

.

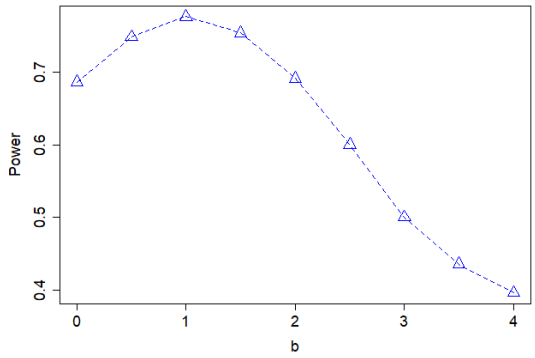



Figure 4: Power of the generalized distance covariance test based on $\mathcal{V}_b$ for the overdominant model in (5.4)

.

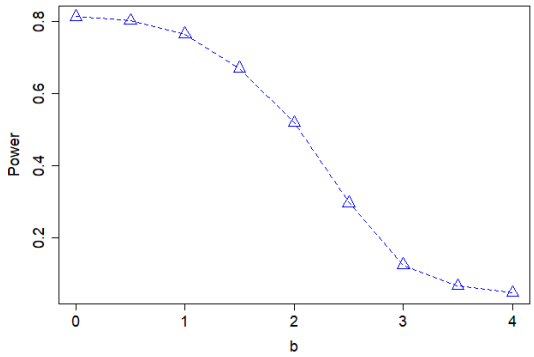



Figure 5: Power of the generalized distance covariance test based on $\mathcal{V}_b$ for the purely heterozygous model in (5.5)

.

## 6. DISCUSSION, CONCLUSIONS AND FUTURE WORK

In this work, we have derived a novel method for testing the association of SNPs with a quantitative response based on the generalized distance covariance $\mathcal{V}_b$. We have further provided a model-based interpretation for the method and investigated which choice of the parameter $b$ is meaningful for applications. Edelmann and Goeman (2021) have derived model-based extensions for the generalized distance covariance that allow e.g. for testing in survival data or generalized models. Therefore, a natural line of future work would be to reformulate GWAS with binary or survival responses in terms of generalized distance covariance.

On the other hand, considering the interpretation of $\mathcal{V}_b$ as test statistics for codominant models in Section 4.2, this paper can be regarded as first step into developing distance covariance models for ordinal data. We also aim at extending the given methodology to ordinal variables $X$ with more than three values. We will also expect to be granted access to SNP data from international consortia before the end of the month, in order to display relevant data examples.

A longer version of this paper, including proofs, data experiments and an extension to deal with nuisance variables (all of which could not be included for the sake of brevity) can be obtained upon request to any of the coauthors.

## REFERENCES

Castro-Prado, F., Costas, J., González-Manteiga, W. and Penas, D. R. (2020+). Searching for genetic interactions in complex disease by using distance correlation. [Preprint.] Available at https://arxiv.org/abs/2012.05285.

Cox D. R. and Hinkley, D. V. (1979) Theoretical Statistics. Chapman and Hall/CRC.

Edelmann, D. and Goeman, J. J. (2021) A regression perspective on generalized distance covariance and the Hilbert–Schmidt independence criterion. [To appear.] Statistical Science.

Gillespie J. H. (2004) Population Genetics: A Concise Guide. The Johns Hopkins University Press.

Goeman, J. J., van de Geer, S. and van Houwelingen, H. (2006) Testing against a high dimensional alternative. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68, 477–493.

Hu, C., Pozdnyakov, V. and Yan, J. (2020) Density and distribution evaluation for convolution of independent gamma variables. Computational Statistics 35, 327–342.

Lettre, G., Lange, C. and Hirschhorn, J. N. (2007) Genetic model testing and statistical power in populationâbased association studies of quantitative traits. Genetic Epidemiology, 31, 358–362.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Annals of Statistics, 41, 2263–2291.

Székely, G. J. and Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. Annals of Statistics, 35, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2009) Brownian distance covariance. Annals of Applied Statistics, 3, 1236–1265.

Visscher, P., Wray, N., Zhang, Q., Sklar, P., McCarthy, M., Brown, M. and Yiang, J. (2017) 10 years of GWAS discovery: Biology, function, and translation. American Journal of Human Genetics, 101, 5–22.

Zhang, Y., Qi, G., Park, J.-H. and Chatterjee, N. (2018) Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nature Genetics, 50, 1318–1326.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

**Resolviendo el problema de distribuir la reducción de las emisiones de $CO_2$ con el paquete de R "ClaimsProblems"**

Iago Núñez Lugilde[1], Miguel Ángel Mirás Calvo[2],
Carmen Quinteiro Sandomingo[2] y Estela Sánchez-Rodríguez[1]

[1]SIDOR. Departamento de Estatística e Investigación Operativa. Universidade de Vigo.
[2]Departamento de Matemáticas. Universidade de Vigo.

**RESUMEN**

Los problemas de reparto, o de bancarrota, surgen cuando se debe realizar una distribución de un bien escaso entre un conjunto finito de demandantes. En noviembre de 2019, el *Programa de las Naciones Unidas para el Medio Ambiente* publicó su décimo informe anual "Emissions Gap Report". Se acordó que, para frenar el cambio climático, la temperatura del planeta debería aumentar menos de 1,5ºC en el año 2030 y, para conseguirlo, las emisiones de $CO_2$ deberían disminuir un 7,6 % anual en el período 2020-2030. Esta situación puede ser tratada desde el punto de vista de los problemas de reparto, considerando que las emisiones totales no deben superar un cierto umbral y repartiendo este entre los diferentes países. En nuestro trabajo analizamos este problema utilizando el paquete de R "ClaimsProblems", obteniendo tanto las asignaciones propuestas por las principales reglas de distribución estudiadas en la literatura como la sugerida por la regla de reparto que consiste en calcular la esperanza matemática cuando consideramos la distribución uniforme sobre el conjunto de repartos estables. De forma complementaria, para ayudar al decisor, comparamos las soluciones establecidas por todas las reglas utilizando medidas de discrepancia, como *el índice de Gini* o *el índice de desviación de la proporcionalidad*, además de curvas que permiten observar la evolución de los repartos a lo largo del tiempo.

**Palabras y frases clave**: Problemas de reparto, reglas de reparto, ranking de Lorenz, algoritmos, emisiones de $CO_2$, cambio climático.

# 1. Introducción

Los problemas de reparto engloban una gran cantidad de situaciones: la quiebra de una empresa, el reparto de una herencia, el reparto del pago de impuestos, la distribución de suministros insuficientes como alimentos o vacunas, o el reparto de las emisiones de $CO_2$ para frenar el cambio climático. Todos estos escenarios tienen una base teórica común: un recurso escaso para repartir y un conjunto de demandantes que solicitan una parte, el total del recurso o incluso un valor mayor. El estudio formal de estos problemas comenzó con los artículos [11] y [1].

Una regla es una forma de repartir ese recurso escaso entre todos los demandantes. Por lo tanto, para cada problema de reparto, una regla debe seleccionar un vector que cumpla tres requisitos básicos: a ningún demandante se le debe pedir que pague, a ningún demandante se le debería conceder más que su propia demanda, y la suma del reparto debe ser igual al recurso. El conjunto de asignaciones que cumplen estos requisitos se conoce como el conjunto de repartos estables. Del gran inventario de reglas, las más conocidas y utilizadas son las siguientes: las reglas proporcional y proporcional ajustada, las reglas de igual ganancia y de igual pérdida, las reglas de llegadas aleatorias, la regla del Talmud, la regla igualitaria restringida, la regla de Piniles, la regla de superposición mínima y la regla de Domínguez-Thomson. En [12] se encuentra una amplia discusión sobre las mismas.

Además de las diez reglas mencionadas, consideraremos la regla de la media de los repartos estables, una nueva regla de división para problemas de reparto, introducida en [9], que selecciona el centro geométrico (centroide) del conjunto de los repartos estables. Desde un punto de vista estadístico, la media de los repartos estables selecciona la esperanza de la distribución uniforme (continua) sobre un conjunto convexo y compacto, el conjunto de los repartos estables. Esta regla coincide con el core-center (véase [4]) del juego cooperativo asociado al problema de reparto.

"ClaimsProblems" (véase [10]) es un paquete de R centrado en problemas de reparto. En él se implementan métodos específicos para cada regla basados en los parámetros que definen un problema de reparto: la dotación inicial y el vector de demandas. En particular, nuestro procedimiento para obtener la regla de llegadas aleatorias se puede aplicar a poblaciones más grandes y es más rápido que los métodos generales utilizados hasta ahora. Para obtener la media de los repartos estables utilizamos el algoritmo descrito en [6].

Además, el paquete incluye funciones gráficas específicas que posibilitan el análisis de problemas de reparto. Obviamente, el conjunto de repartos estables solo se puede dibujar para problemas con un máximo de tres demandantes, o de cuatro si se trabaja en la proyección. Siguiendo los métodos descritos en [12], el paquete proporciona funciones adicionales que permiten representar los repartos elegidos por una regla a medida que la dotación varía desde 0 hasta la suma de las reclamaciones.

La división proporcional se toma a menudo como la definición de equidad para los problemas de reparto. Además de *la curva de Lorenz* y *el índice de Gini* del vector de reparto seleccionado por una regla dada, el paquete permite obtener *el índice de desviación de la proporcionalidad* destinado a cuantificar en qué medida una asignación se aleja del reparto proporcional ([8]). Utilizando "ClaimsProblems" es posible representar también la trayectoria que traza el índice correspondiente en función de la dotación, *index path*. Para comparar reglas en los problemas de reparto es común utilizar el criterio de dominación de Lorenz, ver [2] y [7]. En "ClaimsProblems" se encuentra una función que proporciona dicha información sobre los repartos dados por dos reglas.

Un problema de actualidad donde se encuentran presentes los problemas de reparto consiste en la distribución de las emisiones de $CO_2$ entre todos los países para frenar el cambio climático. Según el *Programa de de las Naciones Unidas para el Medio Ambiente* (2019) (véase [13]) para cumplir el acuerdo de París, las emisiones de $CO_2$ deben disminuir un 7,6 % anual en la próxima década. Con este objetivo, presentamos un modelo dinámico que analiza cómo deben disminuir dichas emisiones en cada país por año basándonos en los repartos propuestos por la regla proporcional, la regla del Talmud, la regla de las llegadas aleatorias, la media de los repartos estables y las reglas de igual ganancia e igual pérdida. Este ejemplo ilustra el comportamiento de la media de los repartos estables y destaca algunas similitudes y discrepancias con las otras reglas, utilizando las herramientas gráficas e índices que hemos implementado.

En la sección 2 del trabajo presentamos la base teórica de los problemas de reparto junto con los resultados resumidos de nuestras investigaciones. En la sección 3 presentamos el paquete "ClaimsProblems", citando algunas de sus funciones y explicando brevemente su uso. Por último, en la sección 4 analizamos el problema del reparto de $CO_2$ con indicaciones que pueden ayudar al decisor que tenga que pautar la forma más efectiva de reducir las emisiones en la próxima década.

## 2. Teoría de los problemas de reparto

Sea $N$ un subconjunto finito de $\mathbb{N}$ y sea $|N| = n$ el número de elementos de $N$. Dados $x \in \mathbb{R}^N$ y $S \in 2^N$, denotaremos $x(S) = \sum_{i \in S} x_i$.

Un problema de reparto con un conjunto de demandantes $N$ es un par $(E, d)$ donde $E > 0$ es el recurso a dividir y $d \in \mathbb{R}^N$ es el vector de demandas que cumple que $d_i \geq 0$ para todo $i \in N$ y $d(N) > E$. Denotamos la clase de problemas de reparto con conjunto de demandantes $N$ por $B^N$. Los subdominios de los problemas de reparto para los que recurso es menor o mayor que la mitad de la suma de las demandas, $B_L^N = \left\{(E, d) \in B^N : E \leq \frac{1}{2} d(N)\right\}$ y $B_H^N = \left\{(E, d) \in B^N : E \geq \frac{1}{2} d(N)\right\}$, se denominan dominio inferior y dominio superior, respectivamente.

El mínimo derecho del demandante $i \in N$ en el problema $(E, d) \in B^N$ es la cantidad $m_i(E, d) = \text{máx}\{0, E - d(N \backslash \{i\})\}$. La demanda truncada del demandante $i \in N$ en el problema $(E, d) \in B^N$ es la cantidad $t_i(E, d) = \text{mín}\{E, d_i\}$. Sean $m(E, d) = \big(m_i(E, d)\big)_{i \in N}$ y $t(E, d) = \big(t_i(E, d)\big)_{i \in N}$.

Un vector $x \in \mathbb{R}^N$ es un reparto estable para el problema $(E, d) \in B^N$ si $0 \leq x_i \leq d_i$ para

todo $i \in N$ y $x(N) = E$. Sea $\chi(E,d)$ el conjunto de repartos estables para $(E,d) \in B^N$. En [12] se demuestra que:

$$\chi(E,d) = \big\{x \in \mathbb{R}^N \colon x(N) = E,\ m_i(E,d) \leq x_i \leq t_i(E,d) \text{ para todo } i \in N\big\}.$$

Para simplificar supondremos que $N = \{1,\ldots,n\}$. El conjunto de repartos estables $\chi(E,d)$ es un politopo convexo no vacío de dimensión menor o igual que $n-1$. Dado un politopo convexo $K \subset \mathbb{R}^N$, denotaremos por $Vol(K)$ su medida de Lebesgue $(n-1)$-dimensional y por $\mu(K)$ su centro de gravedad (centroide).

Una regla es una función $\mathcal{R}\colon B^N \to \mathbb{R}^N$ que asigna a cada problema $(E,d) \in B^N$ un reparto $\mathcal{R}(E,d) \in \chi(E,d)$. Definimos las principales reglas que utilizaremos en este trabajo.

- Regla concede y divide (CD): Para $|N| = 2$ y para cada $(E,d) \in B^N$,

$$\mathrm{CD}(E,d) = \begin{cases} \left(\frac{E}{2}, \frac{E}{2}\right) & \text{si } 0 \leq E \leq d_1 \\ \left(\frac{d_1}{2}, E - \frac{d_1}{2}\right) & \text{si } d_1 \leq E \leq d_2 \\ \left(\frac{E+d_1-d_2}{2}, \frac{E-d_1+d_2}{2}\right) & \text{si } d_2 \leq E \leq d_1 + d_2 \end{cases}$$

- Regla proporcional (PRO): Para cada $(E,d) \in B^N$ y cada $i \in N$, $\mathrm{PRO}_i(E,d) = \frac{d_i}{d(N)}E$.

- Regla de igual ganancia (CEA): Para cada $(E,d) \in B^N$ y cada $i \in N$, $\mathrm{CEA}_i(E,d) = \text{mín}\{\alpha, d_i\}$, donde $\alpha \geq 0$ es tal que $E = \sum_{i \in N} \mathrm{CEA}_i(E,d)$.

- Regla de igual pérdida (CEL): Para cada $(E,d) \in B^N$ y cada $i \in N$, $\mathrm{CEL}_i(E,d) = \text{máx}\{0, d_i - \beta\}$, donde $\beta \geq 0$ es tal que $E = \sum_{i \in N} \mathrm{CEL}_i(E,d)$.

- Regla del Talmud (T): Para cada $(E,d) \in B^N$ y cada $i \in N$,

$$\mathrm{T}_i(E,d) = \begin{cases} \mathrm{CEA}_i(E, \frac{d}{2}) & \text{si } E \leq \frac{1}{2}d(N) \\ d_i - \mathrm{CEA}_i(d(N) - E, \frac{d}{2}) & \text{si } E \geq \frac{1}{2}d(N) \end{cases}$$

- Regla de llegadas aleatorias (RA): Para cada $(E,d) \in B^N$ y cada $i \in N$,

$$\mathrm{RA}_i(E,d) = \frac{1}{|N|!} \sum_{\pi \in \Pi^N} \text{mín}\big\{d_i, \text{máx}\{0, E - d(P_\pi(i))\}\big\},$$

  donde $\Pi^N$ es el conjunto de órdenes de $N$ y $P_\pi(i) = \{j \in N \colon \pi(j) < \pi(i)\}$ para $\pi \in \Pi^N$.

- Media de los repartos estables (AA): Para cada $(E,d) \in B^N$ y cada $i \in N$,

$$\mathrm{AA}_i(E,d) = \frac{1}{Vol(\chi(E,d))} \int_{\chi(E,d)} x_i d\mu.$$

  La media de los repartos estables es el centroide del conjunto de los repartos estables.

Además de estas reglas, también se encuentran implementadas en el paquete "ClaimsProblems" la regla propocional ajustada (APRO), la regla de superposición mínima (MO), la regla de Piniles (PIN), la regla de Dominguez y Thomson (DT) y la regla igualitaria restringida (CE) [1]. La regla concede y divide solo está definida para 2 jugadores y las reglas del Talmud, de llegadas aleatorias, proporcional ajustada, superposición mínima y media de los repartos estables coinciden con ella para $|N| = 2$, por lo que se consideran extensiones de la misma.

Las propiedades que cumplen las reglas antes mencionadas y sus caracterizaciones axiomáticas se pueden encontrar en [12], con nuevas contribuciones para la regla proporcional ajustada y la

[1] Las definiciones de estas reglas se pueden encontrar en [12] o en [8].

regla de superposición mínima en [7]. La media de los repartos estables se analiza en [9]. En la Tabla 1 mostramos algunas de estas propiedades.

| **Propiedades** | PRO | CEA | CEL | T | PIN | CE | APRO | RA | MO | AA |
|---|---|---|---|---|---|---|---|---|---|---|
| Preservación del orden | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Preservación del orden bajo variaciones en las demandas | ✓ | ✓ | ✓ | ✓ | ✓ | No | No | ✓ | ✓ | ✓ |
| Invarianza en las demandas truncadas | No | ✓ | No | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Dualidad | ✓ | No | No | ✓ | No | No | ✓ | ✓ | No | ✓ |
| Propiedad del punto medio | ✓ | No | No | ✓ | ✓ | ✓ | ✓ | ✓ | No | ✓ |
| Consistencia | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | No | No | No | No |
| Mínimos derechos primero | No | No | ✓ | ✓ | No | No | ✓ | ✓ | ✓ | ✓ |
| Límite inferior razonable | No | ✓ | No | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Progresividad en el dominio superior | ✓* | No | ✓* | ✓ | No | No | ✓ | No | ✓ | ✓ |
| Regresividad en el dominio inferior | ✓* | ✓* | No | ✓ | ✓ | ✓ | ✓ | No | No | ✓ |
| Diferenciabilidad respecto al estado | ✓ | No | No | No | No | No | No | No | No | ✓** |

Tabla 1: Reglas y propiedades (* Cumple esa propiedad en todo el dominio. ** Para problemas con $n > 2$.).

El estudio detallado de las propiedades o axiomas que cumple una regla junto con sus caracterizaciones axiomáticas son vitales para poder decidir qué solución se adapta mejor al problema en cuestión. Por ejemplo, una regla regresiva es aquella que trata mejor, en términos de unidad demandada, a los agentes que tienen demandas pequeñas en comparación con los agentes que tienen demandas grandes y una regla progresiva tendrá un comportamiento inverso.

Un juego coalicional es un par $(N, v)$ con un número finito de jugadores $N$ y función característica $v : 2^N \to \mathbb{R}$, cumpliendo que $v(\emptyset) = 0$. Sea $G^N$ el conjunto de juegos coaliciones con conjunto de jugadores $N$. El juego coalicional asociado con el problema de reparto $(E, d) \in B^N$, introducido en [11], es el juego con función característica definida por $v(S) = \max\{0, E - d(N\backslash S)\}$ para cada $S \in 2^N$. El conjunto de repartos estables del problema $(E, d) \in B^N$ coincide con el núcleo del juego asociado, es decir, $\chi(E, d) = C(v) = \{x \in I(v) \colon x(S) \geq v(S) \text{ para todo } S \subset N\}$, siendo $I(v) = \{x \in \mathbb{R}^N \colon x(N) = v(N), x_i \geq v(i) \text{ para todo } i \in N\}$ el conjunto de imputaciones.

Una solución puntual de un juego cooperativo es una función que a cada función característica le asigna un vector de $\mathbb{R}^N$. Existen importantes conexiones entre las principales soluciones del juego cooperativo asociado al problema de bancarrota y las reglas que hemos descrito. Por ejemplo la regla de las llegadas aleatorias coincide con el valor de Shapley del juego cooperativo asociado, al igual que la regla del Talmud coincide con el nucleolus, la regla proporcional con el $\tau$-valor y la media de los repartos estables con el core-center.

Nos centraremos ahora en el cálculo de la media de los repartos estables con el algoritmo descrito en [6], el cual se basa en el recubrimiento del conjunto de imputaciones con el núcleo del juego asociado y los núcleos de los juegos de utopía, que definimos a continuación.

**Definición 1** *Sea $(E, d) \in B^N$ problema de reparto y $T \subset N$. El problema de $T$-utopía $(\tilde{E}_T, \tilde{d}_T) \in B^N$ y el vector $a_T \in \mathbb{R}^N$ vienen dados por:*

| | $d(T) \geq E$ | $d(T) < E$ |
|---|---|---|
| $\tilde{E}_T$ | $E$ | $E - d(T) - \sum_{\ell \in N\backslash T} m_\ell(E, d)$ |
| $\tilde{d}_T$ | $(\tilde{d}_T)_i = \begin{cases} d_i & \text{if } i \in T \\ 0 & \text{if } i \notin T \end{cases}$ | $(\tilde{d}_T)_i = \begin{cases} \tilde{E}_T & \text{si } i \in T \\ d_i - m_i(E, d) & \text{si } i \notin T \end{cases}$ |
| $a_T$ | $a_T(i) = 0, \ i \in N$ | $a_T(i) = \begin{cases} d_i & \text{si } i \in T \\ m_i(E, d) & \text{si } i \notin T \end{cases}$ |

*Sea $\tilde{v}_T \in G^N$ el juego coalicional asociado con el problema de $T$-utopía $(\tilde{E}_T, \tilde{d}_T) \in B^N$. El juego de $T$-utopía se define como $v_T = a_T + \tilde{v}_T \in G^N$.*

El siguiente resultado permite identificar las coaliciones $T$ para las cuales el núcleo del juego de $T$-utopía es de dimensión completa. Dado un problema de reparto $(E, d) \in B^N$, consideramos la familia $\mathcal{F} = \{T \subset N \colon |T| \leq n - 2, \ d(T) < E\}$.

**Proposición 1** *Sea $(E,d) \in B^N$, $T \subset N$, y $v_T \in G^N$ el juego de $T$-utopía . Entonces, $Vol(C(v_T)) > 0$, si y solo si, $T \in \mathcal{F}$.*

Dado un juego asociado a un problema de reparto, su conjunto de imputaciones puede ser recubierto por la unión de los núcleos de los juegos de $T$- utopía que pertenecen a la familia $\mathcal{F}$. Además, la intersección de estos núcleos es de volumen nulo.

**Proposición 2** *Sea $(E,d) \in B^N$, $T \in \mathcal{F}$, y $v_T \in G^N$ el juego de $T$-utopía. Si $T$ es un elemento maximal de $\mathcal{F}$, entonces $C(v_T) = I(v_T)$.*

**Teorema 1** *Sea $(E,d) \in B_L^N$, $v$ su juego cooperativo asociado, para cada $T, R \in \mathcal{F}$ tales que $T \neq R$, sean $v_T, v_R \in G^N$ los juegos de $T$-utopía y $R$-utopía, respectivamente. Entonces $Vol(C(v_T) \cap C(v_R)) = 0$ y $I(v) = \bigcup_{T \in \mathcal{F}} C(v_T)$.*

Aplicamos ahora el Teorema 1 a los problemas de $T$-utopía. Sea $T \in \mathcal{F}$, denotaremos por $\mathcal{F}_T = \{S \in \mathcal{F} \colon S \supset T\}$. $T$ es una coalición maximal de $\mathcal{F}$, si y solo si, $\mathcal{F}_T = \{T\}$.

**Teorema 2** *Sea $(E,d) \in B^N$. Si $T \in \mathcal{F}$, entonces $I(v_T) = \bigcup_{S \in \mathcal{F}_T} C(v_S)$.*

Denotamos los volúmenes del conjunto de imputaciones y del núcleo, escalados por el factor $\alpha = \frac{\sqrt{n}}{(n-1)!}$, por $p_T^I = \frac{1}{\alpha}\,\mathrm{Vol}(I(v_T))$ y $p_T = \frac{1}{\alpha}\,\mathrm{Vol}(C(v_T))$.

Presentamos ahora el esquema para el cálculo de la media de los repartos estables en la Figura 1 aplicando los Teoremas 1 y 2. El algoritmo funciona de la siguiente forma[2]: si $\mathcal{F} = \{\emptyset\}$, $C(v) = I(v)$, y $\mathrm{AA}_i(E,d) = \frac{E}{n}$ para todo $i \in N$. En otro caso para cada coalición maximal $T \in \mathcal{F}$, se tiene que $C(v_T) = I(v_T)$. Por lo tanto, vamos hacia atrás en la cardinalidad de las coaliciones y tendremos que para cada coalición $T \in \mathcal{F}$, o $T$ es maximal o podemos descomponer $I(v_T)$ en función de $C(v_T)$ y los núcleos de los juegos de utopía de las coaliciones maximales obtenidas en el paso anterior.

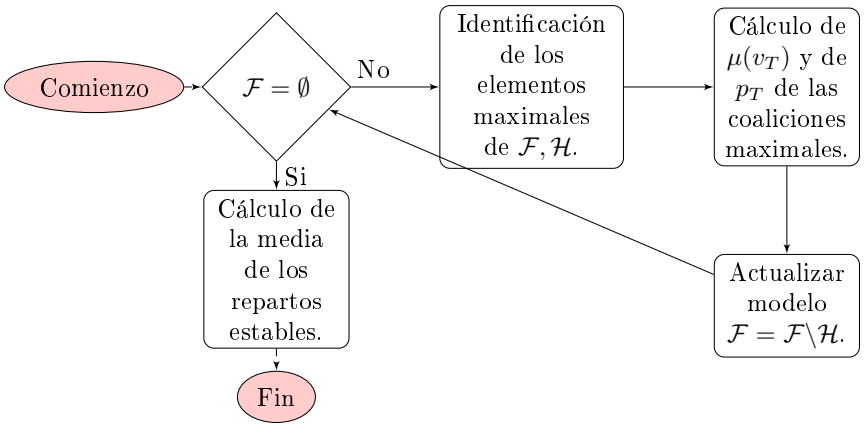



Figura 1: Esquema del algoritmo de la media de los repartos estables.

La media de los repartos estables es dual, por lo que si $E > \frac{d(N)}{2}$, entonces $\mathrm{AA}(E,d) = d - \mathrm{AA}(d(N) - E, d)$. Por lo tanto, podemos restringir el algoritmo a los problemas del subdominio inferior. Además, en [9] se demuestra que, si $E \in [d(N \backslash n), \frac{d(N)}{2}]$, entonces $\mathrm{AA}(E,d) = \frac{d}{2}$. Por lo tanto podemos restringir el uso del algoritmo a los problemas de reparto tales que $E < \min\{\frac{d(N)}{2}, d(N \backslash \{n\})\}$. Sea $\mathcal{I} = \{i \in N : \{i\} \in \mathcal{F}\}$ y $\chi(i) = 1$ si $\{i\} \in \mathcal{F}$ y $\chi(i) = 0$ en otro caso. Simplificamos la notación con $p_i = p_{\{i\}}$ y $p = p_\emptyset$. Usando el procedimiento descrito anteriormente obtenemos una fórmula para el cálculo de la media de los repartos estables.

**Teorema 3** *Sea $(E,d) \in B^N$ tal que $E < \min\{\frac{d(N)}{2}, d(N \backslash \{n\})\}$. Para todo $i \in N$,*

$$\mathrm{AA}_i(E,d) = \tfrac{1}{n}\Big(E + \sum_{j \in \mathcal{I}} \tfrac{p_j}{p} d_j\Big) - \chi(i)\tfrac{p_i}{p} d_i.$$

[2]Sin pérdida de generalidad podemos suponer que el juego es 0-normalizado.

Una vez están definidas todas las reglas, el siguiente paso será tener criterios para discernir que regla es la más apropiada para el problema que queramos resolver. Esta decisión puede apoyarse en el criterio de dominación de Lorenz, el cual se usa comúnmente para comparar reglas en problemas de reparto. Dado un problema $(E,d) \in B^N$, para comparar un par de vectores $x, y \in \chi(E,d)$ con el criterio de Lorenz, primero hay que reordenar las coordenadas de los vectores en orden no decreciente. Hecho esto, se dice que $x$ *Lorenz-domina* $y$ si todas las sumas acumuladas de las coordenadas son mayores para $x$ que para $y$. Dado que todas las reglas que consideramos satisfacen la propiedad de preservación del orden, podemos usar el criterio de dominancia de Lorenz para ver si una regla es más favorable para los demandantes más pequeños en relación con los demandantes más grandes. Sean $\mathcal{R}$ y $\mathcal{R}'$ dos reglas. Decimos que $\mathcal{R}$ *Lorenz-domina* a $\mathcal{R}'$ si $\sum_{i=1}^{k} \mathcal{R}_i(E,d) \geq \sum_{i=1}^{k} \mathcal{R}'_i(E,d)$ para todo $1 \leq k \leq n$ y cada problema $(E,d)$. En general, algunas reglas no son *Lorenz-comparables*, pero al restringir la comparación a los dominios inferior y superior, se obtiene una clasificación de reglas que es más rica. En [2] se proporciona una clasificación de algunas de las reglas centrales, que se completa con los resultados obtenidos en [7].

**Teorema 4** *Sea $S$ el conjunto de reglas que cumplen las propiedades del punto medio, mínimos derechos primero, invarianza en las demandas truncadas, preservación del orden de ganancias (de pérdidas) y regresividad en el dominio inferior (progresividad en el dominio superior), la regla proporcional ajustada es la única regla en $S$ que está Lorenz-dominada por cualquier regla en el dominio inferior (es la única regla en $S$ que Lorenz-domina a cualquier regla en el dominio superior).*

Haciendo uso del Teorema 4 completamos la clasificación de las reglas en función del dominio. Dicho ranking se ilustra en la Figura 2. Una flecha de izquierda a derecha implica *Lorenz-dominancia* y la ausencia de flechas indica que no son comparables. Las líneas verticales indican los diferentes valores que puede tomar $E$, para los que tenemos dos opciones: $d(N\backslash\{n\}) \leq \frac{d(N)}{2}$ (parte izquierda del diagrama) o $d(N\backslash\{n\}) \geq \frac{d(N)}{2}$ (parte derecha del diagrama).

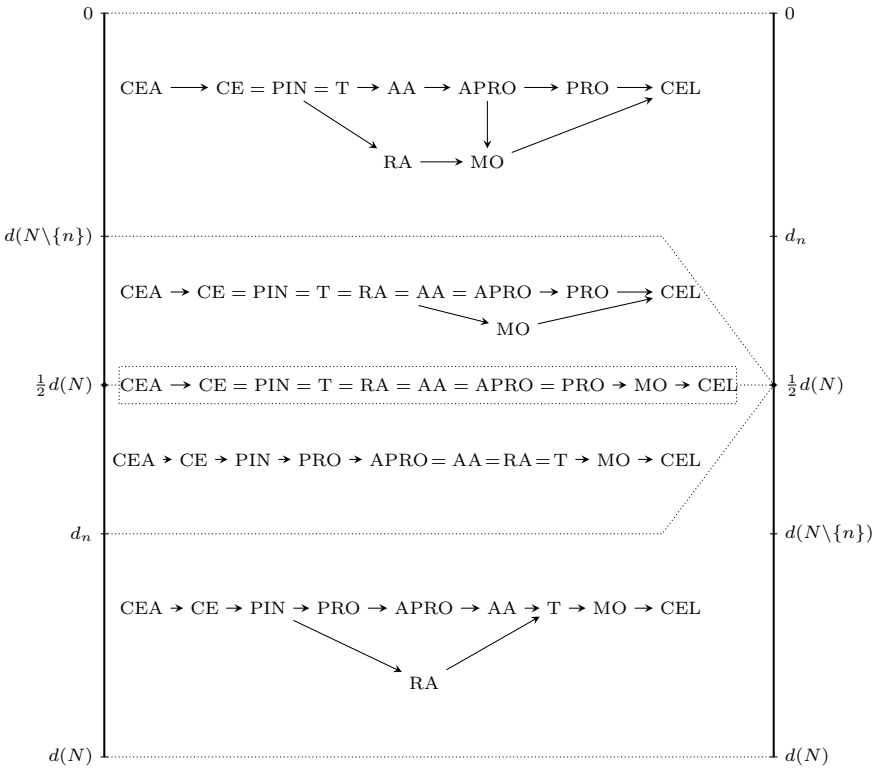



Figura 2: Ranking dinámico de las reglas según el orden de Lorenz.

En 1912, el estadístico y sociólogo Corrado Gini introdujo un coeficiente destinado a medir el grado de desigualdad de ingresos dentro de una población ([3]). Para calcular el coeficiente de Gini, primero hay que encontrar *la curva de Lorenz*, desarrollada por el economista Max O.Lorenz ([5]), que representa en el eje horizontal la proporción de la población, de menor a mayor ingreso, y en el eje vertical el porcentaje acumulado de los ingresos o la riqueza que posee.

Estos conceptos se pueden adaptar a problemas de reparto. Sea $d \in \mathbb{R}^N$ un vector de demandas ordenadas de forma no decreciente, *la curva de Lorenz* de la regla $\mathcal{R}$ para $(E,d) \in B^N$ es el camino poligonal uniendo los $n+1$ puntos $\left(X_k, Y_k\right) = \left(\frac{k}{n}, \frac{1}{E}\sum_{i=1}^{k}\mathcal{R}_i(E,d)\right)$, $k = 0,\ldots,n$, donde $\mathcal{R}_0(E,d) = 0$. *El índice de Gini* de la regla $\mathcal{R}$ para $(E,d) \in B^N$ es la razón del área que se encuentra entre la línea de identidad y *la curva de Lorenz* de la regla sobre el área total debajo de la línea de identidad.

Tomando como referencia estos conceptos, en [8] se define *la curva de demandas y repartos acumulados*, que nos permite comparar la división recomendada por una regla específica con la división proporcional. Dado un problema de reparto $(E,d) \in B^N$ con las demandas ordenadas de forma no decreciente, y una regla $\mathcal{R}$, *la curva de demandas y repartos acumulados* es la trayectoria poligonal que conecta los $n+1$ puntos $\left(U_k, V_k\right) = \left(\frac{1}{d(N)}\sum_{i=0}^{k} d_i, \frac{1}{E}\sum_{i=0}^{k}\mathcal{R}_i(E,d)\right)$, $k = 0,\ldots,n$, donde $d_0 = \mathcal{R}_0 = 0$. Básicamente, esta curva representa el porcentaje de la dotación asignada por la regla a cada proporción acumulada de demandas. Esta curva captura completamente la clasificación de reglas de Lorenz: si una regla $\mathcal{R}$ *Lorenz-domina* una regla $\mathcal{R}'$, entonces, para cada problema de reparto, *la curva de demandas y repartos acumulados* de $\mathcal{R}$ se encuentra por encima de la curva de $\mathcal{R}'$. *El índice de desviación de la proporcionalidad* es la razón del área que se encuentra entre la línea de la proporcionalidad y *la curva de demandas y repartos acumulados*. El índice de la regla $\mathcal{R}$ para un problema $(E,d) \in B^N$ viene dado por:

$$\mathcal{I}(\mathcal{R},E,d) = 1 - \sum_{k=1}^{n}\left(U_k - U_{k-1}\right)\left(V_k + V_{k-1}\right).$$

Fácilmente se puede probar que $-1 \leq I(\mathcal{R},E,d) \leq 1$. Un índice de desviación de proporcionalidad menor que 0 significa que el vector de repartos es más igualitario que el vector de las demandas, es decir, disminuye la desigualdad intrínseca del problema, y un índice de desviación de proporcionalidad mayor que 0 significa que la desigualdad aumenta.

Dada una regla $\mathcal{R}$, la trayectoria del índice, *index path*, para el vector de demandas $d$ es la función que asigna a cada $E \in (0, d(N)]$ *el índice de desviación de proporcionalidad* de la regla $\mathcal{R}$ para el problema $(E,d) \in B^N$ , es decir, $\mathcal{I}_d^{\mathcal{R}}(E) = I(\mathcal{R},E,d)$. La trayectoria del índice es una forma de visualizar, a medida que varía el recurso, la discrepancia de las divisiones dadas por una regla para un vector de demandas, con respecto a la distribución proporcional.

Cuando se trabaja con *el índice de desviación de la proporcionalidad*, se toma de referencia el reparto proporcional. Sin embargo, podría cambiarse la referencia por otra regla. De especial interés es cuantificar el desvío de las reglas repecto a la media de los repartos estables, por ser éste el valor central por excelencia (véase [8]).

## 3. "ClaimsProblems", un paquete para resolver problemas de reparto

El paquete "ClaimsProblems" proporciona funciones fáciles de usar para resolver problemas de reparto. Las funciones incluidas se describen brevemente en las Tablas 2, 3, 4 y 5.

| ***Función y uso*** | ***Descripción*** |
|---|---|
| problemdata(E, d, draw = FALSE) | Vector de mínimos derechos; vector de demandas truncadas; la suma y la semi-suma de las demandas, y un código[3] indicando el dominio al que pertenece el problema.<br>Si draw = TRUE realiza un gráfico de las demandas en orden creciente en el intervalo $[0, d(N)]$. |
| coalitionalgame(E, d, opt = FALSE, lex = FALSE) | Los juegos pesimista y optimista asociados con el problema de reparto. |
| setofawards(E, d, draw = TRUE, col = NULL) | Vértices del conjunto de repartos estables para un problema de reparto con 2, 3, o 4 demandantes.<br>Si draw = TRUE, también realiza la representación gráfica del conjunto de repartos estables. |
| plotrule(E, d, Rule = NULL, awards = NULL, set = TRUE, col = "blue") | Añade en el conjunto de repartos estables un punto asociado con un reparto estable para mostrar la posición del reparto. |

Tabla 2: Funciones que proporcionan información general sobre un problema de reparto.

| ***Función y uso*** | ***Descripción*** |
|---|---|
| AA(E, d, name = FALSE) | Reparto asignado por la media de los repartos estables. |
| APRO(E, d, name = FALSE) | Reparto asignado por la regla proporcional ajustada. |
| CD(E, d, name = FALSE) | Reparto asignado por la regla concede y divide para problemas con 2 demandantes. |
| CE(E, d, name = FALSE) | Reparto asignado por la regla igualitaria restringida. |
| CEA(E, d, name = FALSE) | Reparto asignado por la regla de igual ganancia. |
| CEL(E, d, name = FALSE) | Reparto asignado por la regla de igual pérdida. |
| DT(E, d, name = FALSE) | Reparto asignado por la regla de Dominguez y Thomson. |
| MO(E, d, name = FALSE) | Reparto asignado por la regla de superposición mínima. |
| PIN(E, d, name = FALSE) | Reparto asignado por la regla de Piniles. |
| PRO(E, d, name = FALSE) | Reparto asignado por la regla proporcional. |
| RA(E, d, name = FALSE) | Reparto asignado por la regla de llegadas aleatorias. |
| Talmud(E, d, name = FALSE) | Reparto asignado por la regla del Talmud. |
| allrules(E, d, draw = TRUE) | Data-frame con el vector de repartos seleccionado por todas las reglas.<br>Si draw = TRUE, dibuja un gráfico de barras para cada demandante representando la cantidad que recibe. |

Tabla 3: Resumen de las funciones que resuelven un problema de reparto.

| ***Función y uso*** | ***Descripción*** |
|---|---|
| lorenzdominance(E, d, Rules, Info = TRUE) | El orden de dominancia de Lorenz. |
| lorenzcurve(E, d, Rules, col = NULL, legend = TRUE) | La curva de Lorenz. |
| cumulativecurve(E, d, Rules, col = NULL, legend = TRUE) | La curva de demandas y repartos acumulados. |
| giniindex(E, d, Rule) | El índice de Gini. |
| proportionalityindex(E, d, Rule) | El índice de proporcionalidad. |
| indexpath(d, Rules, col = NULL, points = 201, legend = TRUE) | Trayectoria del índice de proporcionalidad. |

Tabla 4: Resumen de las funciones relacionadas con la comparación de las reglas mediante Lorenz.

[3]cod=1 cuando $(E, d) \in B_L^N$; cod=-1 si $(E, d) \in B_H^N$; cod=0, si $E = \frac{1}{2}d(N)$.

| *Función y uso* | *Descripción* |
|---|---|
| pathawards(d, claimants, Rule, col = red", points = 201) | Representación gráfica del path of awards de una regla para un vector de demandas y un par de demandantes. |
| pathawards3(d, claimants, Rule, col = red", points = 201) | Representación gráfica del path of awards de una regla para un vector de demandas y tres demandantes. |
| schedrule(d, claimants, Rule, col = NULL, points = 201, legend = TRUE) | Schedules of awards de una regla. |
| schedrules(d, claimant, Rules, col = NULL, points = 201, legend = TRUE) | Schedule of awards de varias reglas para un demandante. |
| verticalruleplot(E,d, Rules, col = NULL, legend = TRUE) | Gráfico de barras con los repartos asignados por una regla. |

Tabla 5: Resumen de las funciones gráficas.

## 4. Reparto de las emisiones de $CO_2$

En la *21ª Conferencia de las Partes* (COP 21) que tuvo lugar en París en Diciembre del 2015 se alcanzó un acuerdo histórico para combatir el cambio climático. El objetivo central del Acuerdo de París es: "fortalecer la respuesta global a la amenaza del cambio climático en el contexto del desarrollo sostenible y los esfuerzos para erradicar la pobreza, manteniendo el aumento de la temperatura media mundial por debajo de los 2 grados Celsius y continuar los esfuerzos para limitar el aumento de temperatura a 1,5 grados Celsius, reconociendo que esto reduciría significativamente el riesgo y los impactos del cambio climático". El Acuerdo de París entró en vigor el 4 de noviembre de 2016.

En noviembre de 2019, el *Programa de las Naciones Unidas para el Medio Ambiente* emitió el décimo Informe anual sobre la brecha de emisiones, (véase [13]), donde informa que las emisiones totales alcanzaron un récord de $55,3$ Gt de $CO_2$ en 2018. Para estar en línea con el Acuerdo de París, las emisiones deben bajar un 7,6 por ciento por año desde el 2020 al 2030 para cumplir el objetivo de que la temperatura no aumente más de 1,5°C y 2,7 por ciento por año para cumplir el objetivo de 2°C. En la Figura 3 se muestran las emisiones mundiales de $CO_2$, medidos en gigatones (Gt),[4] desde 1960 hasta 2014 (fuente: Climate Change Data, World Bank Group).

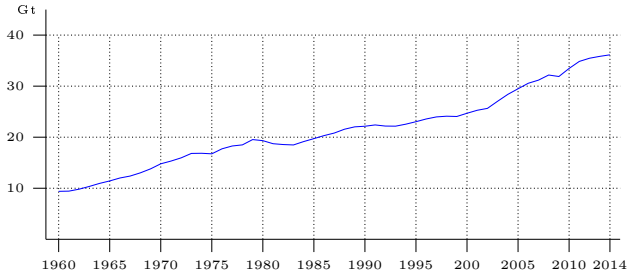



Figura 3: Emisiones de $CO_2$ mundiales desde 1960 a 2014.

Los países emiten cantidades muy diferentes de gases de efecto invernadero a la atmósfera. Incluso si todos ellos estuvieran comprometidos a lograr la meta de 1,5° C, ¿cómo encontrar una reducción razonable de la emisiones de $7,6$ por ciento anual? Estudiemos el problema de reparto donde el recurso es la cantidad de carbono disponible y los demandantes son los países emisores. Siguiendo este modelo, consideramos los 13 países que emitieron la mayor cantidad de dióxido de carbono en 2014: China, EEUU, India, Rusia, Japón, Alemania, Irán, Arabia Saudita, República de Corea, Canadá, Brasil, Sudáfrica y México. Los países restantes están agrupados en 7 regiones geográficas, que son: Unión Europea, resto de Europa, Asia Occidental, resto de Asia, resto de América, resto de África y Oceanía. El año 2014 es el último para el que hay datos de emisiones de $CO_2$ disponibles, tomaremos estos datos como si fueran los del año 2020. La

[4] $1\text{Gt} = 10^6\text{kt} = 10^9\text{t} = 10^{12}\text{kg}$.

Tabla 6 muestra las emisiones de dióxido de carbono estimadas, en kilotones (kt), de las 20 regiones consideradas. Supongamos que cada país se compromete al menos a mantener sus emisiones anuales de $CO_2$ por debajo de la cantidad emitida en 2020. En ese caso, la demanda de cada emisor $d_j$, $j \in N = \{1, \ldots, 20\}$, se corresponde con su emisión estimada en 2014. Naturalmente, denotamos $d = (d_1, \ldots, d_{20})$.

| **1-China** | **2-EEUU** | **3-India** | **4- Resto de la UE** |
|---|---|---|---|
| 10 291 926,878 | 5 225 412,661 | 2 232 729,957 | 2 095 334,801 |
| **5-Resto de Asia** | **6-Rusia** | **7-Asia occidental** | **8-Japón** |
| 1 848 538,367 | 1 736 98,.560 | 1 256 361,871 | 1 206 674,021 |
| **9-Resto de Europa** | **10-Resto de América** | **11-Resto de África** | **12-Alemania** |
| 1 131 240,164 | 919 404,908 | 897 886,952 | 720 363,815 |
| **13-Iran** | **14-Arabia Saudí** | **15-Republica de Korea** | **16-Canadá** |
| 652 392,303 | 601 046,969 | 587 156,373 | 540 614,809 |
| **17-Brasil** | **18-Sudáfrica** | **19-México** | **20-Oceanía** |
| 533 530,165 | 484 495,041 | 481 499,102 | 413 861,287 |

Tabla 6: Emisiones de $CO_2$ (kt) en 2020 de los países/regiones seleccionados.

Según los datos presentados en la Tabla 6, la suma de las emisiones de $CO_2$ en 2020 es $E_0 = 33\,857\,455{,}004$ kt. Ahora, para cumplir el objetivo del Acuerdo de París, las emisiones totales deben caer un 7,6 por ciento cada año. Por lo tanto, para cada $i \in \{1, \ldots, 10\}$, consideramos el problema de reparto $(E_i, d) \in B^N$, donde $E_i = (1 - 0{,}076)^i E_0$.

Para cada problema $(E_i, d) \in B^N$, representamos los repartos que proponen la regla propocional, la regla del Talmud, la media de los repartos estables, la regla de las llegadas aleatorias y las reglas de igual pérdida e igual ganancia.

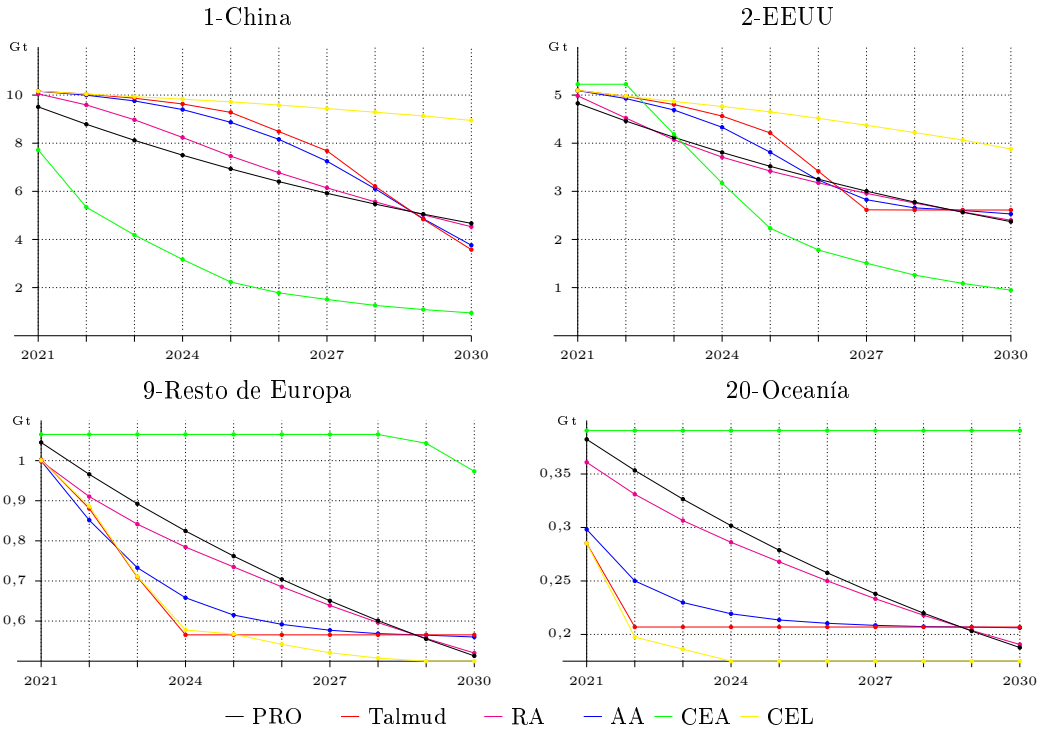



Figura 4: Diferentes patrones de reducción de emisiones dados por las seis reglas.

Las seis reglas nos proporcionan diferentes formas de repartir las emisiones entre los países. En la Figura 4 representamos dichas emisiones para cuatro países o regiones: China, EEUU, resto de

Europa y Oceanía. Elegimos estas regiones debido a que China y EEUU son las regiones que más contaminan, el resto de Europa ocupa una posición intermedia y Oceanía es la región que menos contamina. Así que con estos cuatro casos podemos estudiar el comportamiento de todas las reglas según sus demandas. Mostramos la evolución de la reducción de las emisiones, desde 2021 hasta 2030, recomendadas por las seis reglas.

Observamos patrones muy claros. La regla de igual ganancia *Lorenz-domina* al resto de reglas, por lo que es la que mejor se comporta con los países menos contaminantes y es la que peor se comporta con los más contaminantes, China y EEUU. Nótese que a China le exige reducir sus emisiones en el año 2030 por debajo de un millón de kilotones, algo que parece completamente inverosímil para un país tan grande e industrial, mientras que a Oceanía prácticamente no le exige esfuerzos en los 10 años, por lo que podemos descartar el reparto propuesto por esta regla.

La regla de igual pérdida está *Lorenz-dominada* por el resto de reglas, por lo que es la que menos le exige a los países más contaminantes, mientras que regiones como Oceanía deberían dejar de emitir $CO_2$ a partir del año 2024, otro comportamiento difícil de llevar a la práctica.

Centrémonos en las otras 4 reglas de la Figura 4. Para $i \in \{1, \dots, 8\}$, el recurso $E_i$ es mayor que la mitad de la suma de las demandas, mientras que $E_{10} < E_9 < \frac{1}{2}d(N)$. Por lo tanto, en los 8 primeros años las relaciones de *Lorenz-dominancia* son las mostradas en el dominio superior de la Figura 2, cambiando las mismas en el año 2028.

En un problema de reparto perteneciente al dominio superior, la regla proporcional y la regla de las llegadas aleatorias *Lorenz-dominan* a la regla del Talmud, mientras que la proporcional *Lorenz-domina* a la media de los repartos estables. Por lo tanto, PRO y RA tienen un mejor comportamiento con los pequeños emisores en los 8 primeros años. Sin embargo, en el año 2028 al cambiar el dominio del problema esta relación se invierte y son las reglas del Talmud y la media de los repartos estables quienes ofrecen reducciones menores a los pequeños emisores.

Analizando las propiedades que verifican las reglas (como las concavidades/convexidades y otras más débiles como las visibilidades o la regresividad y progresividad) tendremos criterios en los que sustentar la elección de la más apropiada. Convencer a los países de la viabilidad de una propuesta es fundamental para lograr el compromiso de todos. La media de los repartos estables y el Talmud son progresivas en el dominio superior, por lo que su comportamiento es más favorable con los grandes emisores, mientras que son regresivas en el inferior. La regla de las llegadas aleatorias no es progresiva ni regresiva. Como la regla proporcional es progresiva y regresiva en ambos dominios ocupa una posición intermedia, sin beneficiar más o menos a ningún tipo de demandante, sin embargo no verifica propiedades como invarianza en las demandas truncadas, mínimos derechos primero o los límites inferiores razonables.

Por otra parte, *el índice de desviación de la proporcionalidad* nos indicará como varía la desigualdad en las emisiones iniciales con respecto a los repartos finales. Obviamente la regla proporcional tendrá un índice igual a 0 en cualquier reparto. La regla de llegadas aleatorias no es *Lorenz-comparable* con la regla proporcional en ningún dominio, por lo que no podemos asegurar que aumente o disminuya dicha desigualdad. Sin embargo, podemos calcular el índice para este ejemplo concreto y nos encontramos que, en el primer año, $\mathcal{I}(RA, E_1, d) = 0{,}02518$, es decir, aumenta la desigualdad, pero en último año $\mathcal{I}(RA, E_{10}, d) = -0{,}009$, por lo que la desigualdad disminuye. Para las reglas del Talmud y la media de los repartos estables, que están *Lorenz-dominadas* por la regla proporcional en el dominio superior, en los primeros 8 años la desigualdad aumentará, obteniendo el resultado inverso en los últimos años. Es interesante cuantificar como varía esa desigualdad para poder comparar las reglas. $\mathcal{I}(T, E_1, d) = 0{,}04229$, $\mathcal{I}(AA, E_1, d) = 0{,}04148$, $\mathcal{I}(T, E_{10}, d) = -0{,}07$ y $\mathcal{I}(AA, E_{10}, d) = -0{,}062$. Nótese que en valor absoluto estos valores son superiores a los que obtuvimos con la media de los repartos estables, por lo que la variación en la desigualdad es mayor. Repartiendo con estas dos últimas reglas, en el año 2030 las emisiones de $CO_2$ serían más igualitarias.

Además, exigirles menos a los países más contaminantes en los primeros años, sabiendo que la tendencia se invertirá, puede ser interesante debido a estos son más reacios a entrar en estos acuerdos. Dicho reparto podría llevar a un mayor compromiso que repercuta en conseguir mantener las tasas de emisiones de $CO_2$ bajo control, permitiendo el desarrollo de nuevas formas de producción para frenar la polución.

Por último, es posible pensar en una versión continua de nuestro modelo considerando para cada $t \in [0, 10]$ el problema de reparto $(E_t, d) \in B^N$, donde $E_t = E_0 e^{-0{,}076t}$. La ruta seguida por

el vector de repartos elegido por cualquiera de las seis reglas, digamos $\mathcal{R}$, a medida que el tiempo aumenta de 0 a 10, es decir, la función $\mathcal{R}(t) = \mathcal{R}(E_t, d)$ es una estrategia dinámica de reducción de emisiones. Ahora, una regla $\mathcal{R}$ satisface la propiedad de diferenciabilidad si $\mathcal{R}(\cdot, d)$ es una función diferenciable. La regla proporcional y la media de los repartos estables cuando $|N| > 2$ (véase [9]) son las únicas que satisfacen esta propiedad. Como consecuencia, la trayectoria de reducción de emisiones correspondiente a la regla proporcional y a la media de los repartos estables no presentan cambios bruscos. Los gráficos que se muestran en la Figura 4 corresponden al enfoque de tiempo discreto, pero se puede observar fácilmente que las reglas del Talmud y de las llegadas aleatorias no varían suavemente.

Si consideramos el problema en tiempo continuo, el reparto del Talmud y la media de los repartos estables parecen los más deseables, teniendo esta última regla una ventaja adicional, la diferenciabilidad, una propiedad añadida que provoca variaciones suaves al considerar un modelo dinámico.

**Agradecimientos**

Este trabajo fue financiado por FEDER/Ministerio de Ciencia, Innovación y Universidades. Agencia Estatal de Investigación MTM2017-87197-C3-2-P y PID2019-106281GB-I00 (AEI/FEDER,UE).

# Referencias

[1] R. J. Aumann and M. Maschler. Game theoretic analysis of a bankruptcy problem from the talmud. *Journal of Economic Theory*, 36:195–213, 1985.

[2] K. Bosmans and L. Lauwers. Lorenz comparisons of nine rules for the adjudication of conflicting claims. *International Journal of Game Theory*, 40:791–807, 2011.

[3] C. Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[4] J. González Díaz and E. Sánchez Rodríguez. A natural selection from the core of a tu game: the core-center. *International Journal of Game Theory*, 36:27–46, 2007.

[5] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9:209–219, 1905.

[6] M. A. Mirás Calvo, I. Núñez Lugilde, C. Quinteiro Sandomingo, and E. Sánchez Rodríguez. An algorithm to compute the core-center rule of a claims problem with an application to the allocation of $CO_2$ emissions. *Preprint*, 2020.

[7] M. A. Mirás Calvo, I. Núñez Lugilde, C. Quinteiro Sandomingo, and E. Sánchez Rodríguez. The adjusted proportional and the minimal overlap rules restricted to the lower-half, higher-half, and middle domains. *Ecobas*, 2021.

[8] M. A. Mirás Calvo, I. Núñez Lugilde, C. Quinteiro Sandomingo, and E. Sánchez Rodríguez. Deviation from proportionality and lorenz-dominance between the average of awards and the standard rules for claims problems. *Ecobas*, 2021.

[9] M. A. Mirás Calvo, C. Quinteiro Sandomingo, and E. Sánchez Rodríguez. The core-center rule for the bankruptcy problem. *Ecobas*, 2020.

[10] I. Núñez Lugilde, M. A. Mirás Calvo, C. Quinteiro Sandomingo, and E. Sánchez Rodríguez. *ClaimsProblems: Analysis of Conflicting Claims*, 2021. R package version 0.1.0. `https://CRAN.R-project.org/package=ClaimsProblems`.

[11] B. O'Neill. A problem of rights arbitration from the talmud. *Mathematical Social Sciences*, 2:345–371, 1982.

[12] W. Thomson. How to divide when there isn't enough. *Econometric Society Monographs. Cambridge University Press*, 2019.

[13] United Nations Environment Program. *Emissions Gap Report 2019*. UNEP, Nairobi, 2019.

# Pósters

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ESTUDIO Y SIMULACIÓN DE MEDIDAS FRENTE A LA COVID-19 EN LA ENM

Álvarez Hernández, M.[1], Díaz Amado, A.[2] y González-Cela Echevarría, G.[1]

[1] Centro Universitario de la Defensa – Escuela Naval Militar.

[2] Armada Española.

## RESUMEN

La aparición del coronavirus SARS-CoV-2 a finales de 2019 provocó el desconcierto, incluso para la comunidad científica, al convertirse de forma acelerada en epidemia mundial. Esto llevó al estudio exhaustivo de las formas de propagación y de las medidas higiénico-sanitarias necesarias para evitar el contagio y el consecuente desarrollo de la enfermedad asociada (COVID-19). Muestra de ello son las distintas normativas que a lo largo del pasado curso 2020-2021 se implantaron en las facultades y centros de enseñanza, y que permitieron el desarrollo "normal" de la formación de los/as estudiantes.

En un recinto, como es la Escuela Naval Militar (ENM), donde el alumnado convive diariamente mientras recibe su formación universitaria y castrense, se han establecido los protocolos y restricciones pertinentes en la lucha contra esta pandemia. De ahí que, el objetivo de este trabajo sea analizar la influencia que han tenido y tienen estas medidas anti-COVID-19, mostrando su relevancia y efectividad en un recinto como es la ENM.

Para el propósito de esta investigación, se comenzó con el estudio de los factores que pueden afectar a la expansión del coronavirus, puntualizando cuál o cuáles influyen más en la transmisión dentro del ámbito específico de una escuela militar en régimen de internado. Una vez determinados dichos factores, se planteó un diseño factorial completo $2^k$ para establecer qué variables e interacciones tienen mayor relevancia en las distintas situaciones de la vida diaria de un/a estudiante de la ENM. Utilizando uno de los modelos epidemiológicos clásicos de propagación de enfermedades, como es el modelo SIR, se realizaron varias simulaciones con el software Anylogic para cada condición de estudio (tratamiento), mostrando la propagación a lo largo del tiempo y el número de individuos que han pasado la enfermedad a los 30 días del inicio del brote.

Los resultados muestran el comportamiento temporal de la dispersión del virus en cada una de los escenarios planteados. A través del estudio de un modelo lineal generalizado se determina la significación estadística de cada tratamiento, verificándose el alto porcentaje de explicación del modelo respecto al número de contagios en el caso del uso de las distintas pautas establecidas por los protocolos.

Con todo ello, se demuestra la importancia de las medidas anti-COVID-19, siendo especialmente relevante el uso de la mascarilla. Se observa cómo en caso de la propagación de la enfermedad, ésta impide un rápido contagio entre el alumnado, evitando un colapso de los servicios sanitarios y de los lugares de residencia para infectados.

**Palabras y frases clave:** Diseño de Experimentos, Modelo SIR, COVID-19.

## REFERENCIAS

Burgos Simón, C., Cortés, J. C., López Navarro, E., Martínez-Rodríguez, D., Martínez-Rodríguez, P., Julián, R. S. y Villanueva, R. (2020). Modelo para el estudio de la dinámica de transmisión del SARS-COV-2 y la enfermedad del COVID19. Descripción técnica. Instituto Universitario de Matemática Multidisciplinar - Universidad Politécnica de Valencia. Disponible en https://covid19.webs.upv.es/INFORMES/Explicacion_tecnica.pdf

Díaz Amado, A., González-Cela Echevarría, G. y Álvarez Hernández, M. (2021). Estudio y simulación de impacto del COVID-19 en la ENM. Trabajo de fin de grado en Grado de Ingeniería Mecánica. Universidad de Vigo. Disponible en http://calderon.cud.uvigo.es/handle/123456789/418

Godio, A., Pace, F. and Vergnano, A. (2020). SEIR modeling of the italian epidemic of SARS-CoV-2 using computational swarm intelligence. International Journal of Environmental Research and Public Health, 17 (10), 3535.

Gutiérrez Pulido, H. y De la Vara Salazar, R. (2008). Análisis y diseño de experimentos (2º edición), 46 (5). México D.F. McGraw-Hill.

Gutiérrez, J. M. y Varona, J. L. (2020). Análisis de la posible evolución de la epidemia de coronavirus COVID-19 por medio de un modelo SEIR. Departamento de Matemáticas y Computación - Universidad de La Rioja. Disponible en https://www.unirioja.es/apnoticias/servlet/Archivo?C_BINARIO=12051

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A, 115, 700-721.

Lelieveld, J., Helleis, F., Borrmann, S., Cheng, Y., Drewnick, F., Haug, G., Klimach, T., Sciare, J., Su, H. and Pöschl, U. (2020). Model Calculations of Aerosol Transmission and Infection Risk of COVID-19 in Indoor Environments. International Journal of Environmental Research and Public Health, 17 (21), 8114.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# ALGORITMO HEURÍSTICO EN DÚAS ETAPAS PARA UNHA VARIANTE DO PROBLEMA DE RUTAS DE VEHÍCULOS COMPARTIMENTADOS CON DEMANDAS ESTOCÁSTICAS

Juan Carlos Gonçalves-Dosantos[1] e Balbina Virginia Casas-Méndez[2]

[1] CITIC, Centro TIC de Investigación, MODES, Grupo de Modelización, Optimización e Inferencia Estatística, Departamento de Matemáticas, Universidade da Coruña, España; juan.carlos.goncalves@udc.es

[2] IMAT, Instituto de Matemáticas, MODESTYA, Grupo de Modelos de Optimización, Decisión, Estatística e Aplicacións, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, España; balbina.casas.mendez@usc.es

## RESUMO

Neste traballo deséñase o modelo de programación lineal enteira mixta dun problema de rutas de vehículos, no que as demandas dos clientes considéranse estocásticas e os vehículos divídense en compartimentos. O problema está motivado polas necesidades dalgunhas cooperativas agrícolas españolas, que fabrican diversos tipos de alimentos para o gando. Os vehículos e os seus compartimentos teñen diferentes capacidades onde cada compartimento só pode conter un tipo de alimento. Ademais, o acceso a cada granxa pode estar restrinxido só a determinados vehículos. Primeiramente, estúdase o alcance da resolución exacta do modelo, a cal se consegue nun tempo razoable unicamente en problemas cun pequeno número de clientes. Para problemas de tamaño mediano ou grande, proponse a resolución do modelo mediante un algoritmo de dous pasos. Se parte dunha heurística construtiva, seguida dunha fase de mellora que utiliza a lóxica da búsqueda tabu. Aplicamos este algoritmo a un conxunto de datos históricos. Finalmente, lévase a cabo un estudo de simulación, baseado na creación dun conxunto de problemas de referencia resultante da adaptación ao noso problema de *benchmarks* da literatura. En ambos os casos, presentamos os resultados baseados en diferentes valores dos parámetros do modelo.

**Palabras e frases chave:** Problemas de rutas de vehículos; demandas estocásticas; vehículos con múltiples compartimentos; accesibilidade limitada; cooperativas agrícolas; algoritmos heurísticos.

## 1. INTRODUCIÓN

Os problemas de rutas de vehículos (VRP) contemplan o deseño dun conxunto de rutas de custo mínimo para unha frota de vehículos co fin de atender as demandas dun grupo de clientes. A partir deste problema básico xorden outros máis complexos, engadindo diferentes tipos de restricións de forma que se adapten á realidade o máis posible. Entre eles, estudaremos os problemas de rutas de vehículos con múltiples compartimentos (MC-VRP). Estes problemas comparten o mesmo obxectivo que os VRP, pero teñen produtos diferentes (incompatibles), que deben ser transportados en compartimentos de vehículos independentes. Moitos destes problemas, con todo, non teñen en conta un aspecto importante: non consideran a incerteza, isto é, o feito de que os parámetros que constitúen o problema na vida real son aleatorios, o que non se soluciona nin sequera utilizando unha cantidade adicional de restricións. Cando algún dos parámetros do modelo é aleatorio, atopámonos ante un problema estocástico de rutas de vehículos (SVRP). En contraste cos modelos deterministas, estes teñen unha literatura bastante dispersa e desordenada. Dos traballos existentes na literatura, un dos problemas máis estudados é cando consideramos as demandas como variables aleatorias. Son os chamados problemas de rutas de vehículos con demandas estocásticas (VRPSD). Do mesmo xeito que no caso determinista, queremos determinar un conxunto fixo de rutas coa mínima distancia total, que inclúe a distancia percorrida no conxunto fixo de rutas e a distancia extra esperada (chamámola distancia esperada) que pode requirir unha realización

particular das variables aleatorias. Estas distancias extra débense ao feito de que a demanda nas rutas pode superar ocasionalmente a capacidade do vehículo. Dado que a demanda é aleatoria, isto coñécese gradualmente a medida que o vehículo completa a súa ruta, o que lle obriga a regresar ao depósito antes de continuar a ruta. Este traballo combina os dous aspectos de rutas de vehículos mencionados anteriormente: os vehículos multicompartimentados e as demandas estocásticas. Denominámolo problema de rutas de vehículos multicompartimentados con demandas estocásticas (MC-VRPSD). Non hai moitos traballos na literatura que analicen este dobre problema. Ademais, incorporamos as seguintes restricións: nun compartimento non podemos mesturar diferentes produtos ou produtos demandados por diferentes clientes, pode haber pedidos urxentes (é unha restrición denominada xanela de tempo), e pode haber problemas de accesibilidade dos camións a determinados clientes. Todos estes ingredientes fan que o noso modelo sexa novo, ata onde sabemos. A motivación deste traballo é un problema tomado da vida real. Para resolver o problema, deseñamos un algoritmo construtivo, estendendo o coñecido algoritmo de Clarke e Wright (1964). A nosa proposta proporciona unha solución inicial ao problema que ten en conta as restricións que impiden a mestura de produtos na mesma tolva dun camión e as urxencias dos pedidos cando se define o aforro asociado á inserción dun novo cliente nunha ruta. Ademais, tendo en conta a natureza estocástica do problema e a posibilidade de grandes demandas en relación coa capacidade das tolvas, permitiremos que a demanda dun cliente poida ser dividida e servida en diferentes tolvas do mesmo camión ou en diferentes camións. A solución final que propoñemos é unha mellora desta, obtida mediante un enfoque metaheurístico baseado nunha búsqueda tabú. Os algoritmos están programados na linguaxe *R*. Tamén é interesante destacar que os algoritmos deseñados perseguen o dobre obxectivo de maximizar a carga transportada e minimizar os custos de transporte. Probamos o algoritmo con datos históricos reais e tamén creamos un conxunto de casos de proba simulados, sobre os que se realizou un estudo computacional. En ambos casos, presentamos os resultados baseados en diferentes valores dos parámetros do modelo. Os resultados mostran a robustez do algoritmo e que este proporciona bos resultados en curtos períodos de tempo.

## 2. REVISIÓN BIBLIOGRÁFICA

Nesta sección, realizamos unha ampla revisión dos traballos relacionados co problema do noso interese, centrándonos especialmente nos problemas de rutas de vehículos compartimentados. Moitos estudos anteriores centráronse na resolución de VRP deterministas. Por exemplo, en Dantzig e Ramser (1959), propuxéronse varios problemas cun obxectivo común -deseñar rutas óptimas para camións de reparto de gasolina entre unha terminal e un gran número de estacións-. Ademais, presentáronse diferentes variantes de VRP e métodos de solución nos estudos de Cordeau et al. (2007), Laporte (2009), e Pollaris et al. (2015). Ademais, houbo algúns traballos sobre SVRP e as súas aplicacións. En Gendreau et al. (1996), Berhan et al. (2014), Dror (2016), Oyola et al. (2017), e Oyola et al. (2018), presentáronse estudos sobre variantes de SVRP e métodos de solución. Desde o punto de vista das aplicacións, Dror et al. (1985) consideraron a entrega de aceite de calefacción doméstica, onde o consumo diario do cliente tamén era aleatorio. En Bertsimas e Simchi-Levi (1996), discutíronse moitas outras aplicacións, como a distribución de cervexa, gasolina ou produtos farmacéuticos. Un caso especial de VRP é aquel en o que os vehículos se dividen en multi-compartimentos (MC-VRP). Brown e Graves (1981) e Brown et al. (1987) consideraron traballos seminales con este tipo de problemas. Consideraron, por separado, o deseño das rutas e a asignación dos produtos aos compartimentos, e así, fixeron uso dunha colección de problemas de viaxantes de comercio. Outro dos primeiros artigos sobre MC-VRP foi o de van der Bruggen et al. (1995). Nel se modelaba a redistribución de produtos nunha gran empresa petroleira e propoñíase un algoritmo heurístico para a súa resolución. Cando se aplicou aos datos da empresa, este algoritmo xerou un aforro de custos con respecto á distribución inicial da empresa. Desenvolvéronse varias aproximacións numéricas para resolver MC-VRPs baixo diferentes restricións. En Chajakis e Guignard (2003), propúxose un algoritmo heurístico baseado na relaxación de Lagrange para a subministración de tendas de conveniencia. En Avella et al. (2004), propúxose unha formulación de partición de conxuntos, na que consideraron un algoritmo exacto de rama e prezo e, finalmente, fixeron uso dun algoritmo heurístico de empaquetamiento/rutas para problemas grandes. No Fallahi et al., comparouse un algoritmo construtivo, un algoritmo memético e unha búsqueda tabú, onde a asignación dos tipos de produto aos compartimentos era fixa. Concluíron que a búsqueda tabú proporcionaba resultados lixeiramente mellores, aínda que requiría máis tempo de cálculo. En Oppen e Lokketangen (2008) propúxose un algoritmo de procura tabú para un problema de transporte de animais en vehículos multicompartimentados con restricións de inventario. En Caramia e Guerriero (2010) investigouse un problema de rutas de vehículos multicompartimento, no que como máximo se podía asignar un tipo de

produto a un compartimento e coa restrición adicional de que algunhas granxas eran pequenas e inaccesibles para os vehículos grandes. Non se proporcionou un modelo que abarcase todo o problema. No seu lugar, utilizouse unha heurística en dúas etapas. Na primeira etapa minimizábase o número de vehículos e asignábanse os clientes aos vehículos. Na segunda, minimizábase o custo das rutas. A restrición de inaccesibilidade, neste caso, manexouse asumindo a existencia de vehículos formados por un camión e un remolque, de maneira que algunhas leiras non podían ser visitadas polo remolque; nese caso, eran visitadas só polo camión. Recentemente, houbo numerosos estudos de MC-VRP que inclúen varias restricións, entre elas os custos de transporte. Outro traballo recente é o de Coelho e Laporte (2015), que definiron e compararon catro categorías do problema de entrega de combustible en varios compartimentos que implican tanto custos de transporte como de inventario. Propuxeron formulacións para cada caso e describiron un algoritmo de rama e corte que resolveu os casos dun e varios períodos, que contiñan ata 50 e 20 clientes, respectivamente. En Archetti et al. (2016), estudouse o problema da entrega de varios produtos, e compararónse os custos de transporte, nos que se utilizan vehículos para unha ou varias mercadorías. Elbek e Wohlk (2016) consideran un problema de recollida diaria de vidro e papel reciclable mediante un vehículo que transporta dous contedores. Estudan como programar a recollida e o transporte para minimizar o custo da operación. Desenvolven un algoritmo de búsqueda de veciñanza variable adaptable para resolver o problema. En comparación co caso real, o algoritmo mostra un uso máis eficiente das capacidades. Silvestrin e Ritt (2017) presentaron unha búsqueda tabú para resolver un MC-VRP. En Henke et al. (2019), considerouse unha variante do MC-VRP, que se deu no contexto da reciclaxe de residuos de vidro. Supúxose que, para a recollida do contido dos envases de vidro, dispoñíase dunha frota de vehículos homoxénea. Para cada vehículo, a capacidade podía separarse discretamente nun número limitado de compartimentos, aos que se asignaban diferentes tipos de residuos de vidro. O obxectivo do problema era minimizar a distancia total que debían percorrer os vehículos de eliminación. Para resolver este problema de forma óptima, desenvolveron e implementaron un algoritmo de ramificación e corte. En Ostermeier e Hübner (2018), considerouse a selección de vehículos para un MC-VRP con compartimentos flexibles. O obxectivo da investigación era mostrar os beneficios de considerar tanto os vehículos dun como de varios compartimentos para a distribución, tendo en conta o custo que supón o uso dos correspondentes tipos de vehículos. O problema resolveuse cunha búsqueda de grandes veciñanzas. Do mesmo xeito que o MC-VRP, iniciáronse algúns traballos para estudar as súas variantes estocásticas. En contraste co VRPSD, a investigación do MC-VRPSD é escasa. En Tatarakis e Minis (2009), abordouse unha variante do problema dunha única ruta, na que a secuencia de clientes (ruta) estaba fixada de antemán. O problema consistía en seleccionar os puntos óptimos de reposición ao longo da ruta. Os autores propuxeron un conxunto de algoritmos de programación dinámica e resolveron os problemas de optimalidad para ata 15 clientes. En Mendoza et al. (2011), propúxose un conxunto de heurísticas de construción, que comprendían versións estocásticas da heurística do veciño máis próximo, da mellor inserción e de Clarke e Wright (Dror e Trudeau, 1986), e estendéronse adicionalmente ao caso multicompartimental. Esta última resultou ser a máis competitiva, segundo os extensos experimentos computacionales realizados en diferentes tipos de instancias. En Pandelis et al. (2012), considerouse o problema de deseñar a traxectoria dun vehículo cun só compartimento, que entrega diferentes demandas de produtos que son variables aleatorias. Utilizaron un algoritmo de programación dinámica para satisfacer as demandas dos clientes co mínimo custo total esperado. En Huang (2015), propúxose un modelo de programación matemática para un MC-SVRP. Consideráronse dous tipos de clientes: os que solicitaban un determinado produto, ou os que debían recoller un produto. Utilizaron dúas frotas de vehículos e formaron dous conxuntos de rutas diferentes. Para a resolución do modelo, utilizaron unha búsqueda tabú que partía dunha solución inicial creada de forma aleatoria. A eficacia do algoritmo probouse mediante diferentes *benchmarks*. En Goodson (2015), propúxose un algoritmo de recocido simulado para un MC-SVRP.

A determinación das rutas óptimas dos vehículos ten unha complexidade *NP- hard*. Os diferentes algoritmos que proporcionan unha solución exacta son interesantes para probar o modelo pero, en xeral, só son útiles cando se considera un pequeno número de clientes. En lugar de solucións exactas, adóitanse considerar as solucións obtidas por algoritmos meta-heurísticos, combinados con heurísticas que proporcionan unha boa solución inicial. Como vimos na panorámica anterior, entre as meta-heurísticas, unha das máis utilizadas é o algoritmo tabu (cf. Glover, 1989, Glover, 1990, e Xia et al., 2018, entre outros), que se caracteriza pola súa simplicidade, rapidez, precisión e robustez. En canto ás heurísticas, os estudos pioneiros en MC-SVRP deseñaron adaptacións do algoritmo de Clarke e Wright (1964).

Para terminar esta revisión, respecto ao problema concreto da distribución de alimentos, os traballos recentes son: Hsu et al. (2007), que consideraron un VRPSD; Ambrosino e Sciomachen (2007) e

Kuznietsov et al. (2016), que estudaron VRPs; e Ruiz et al. (2004), Kandiller et al. (2015), e Guitián de Frutos e Casas-Méndez (2019) que modelaron un MC-VRP.

## 3. O PROBLEMA REAL

Algúns tipos de cooperativas agrícolas producen diferentes tipos de alimentos para animais de granxa, que se distribúen a un gran número de agricultores. Moitas veces, os consumidores están dispersos nunha ampla zona. Tradicionalmente, piden diferentes tipos de alimentos, e a cooperativa ten un prazo de entrega determinado, en función da urxencia do pedido. Esta urxencia está motivada polo feito de que a cantidade de penso que queda ao cliente é pequena e necesita encher rapidamente o seu depósito para garantir a alimentación dos animais. Se o pedido é urxente, debe entregarse nun día, pero os custos de entrega son maiores. Este procedemento pode producir saturacións no sistema, polo que non é ideal para un deseño eficiente das rutas. Consideramos unha cooperativa de Galicia, onde se producen catro tipos de pensos. Nesta cooperativa hai máis de 1.500 clientes e cada un deles realiza como máximo dous pedidos ao mes, normalmente dun só tipo de penso. Esta cooperativa conta cunha frota de diferentes camións. Cada un deles ten varias tolvas/compartimentos, con diferentes capacidades, e cada tolva só pode transportar un tipo de penso. Os camións tamén están restrinxidos a unha cantidade limitada de distancia percorrida por día e carga. O condutor do camión cobra pola distancia percorrida (que tamén pode expresarse en termos de tempo de viaxe) e a carga transportada. Por último, o acceso a cada explotación está restrinxido a uns poucos vehículos. O obxectivo deste traballo é dotar á empresa dunha ferramenta de anticipación, é dicir, dun método para deseñar as súas rutas para futuras entregas (sen necesidade de esperar aos pedidos dos clientes), o que pode mellorar a eficiencia da distribución, reducir os custos asociados e, sobre todo, evitar a saturación do sistema. Coñecendo o número de días transcorridos desde a última entrega e o consumo diario estimado dunha explotación, podemos saber con que urxencia hai que visitala. Ademais, coñecendo os pedidos anteriores podemos estimar unha distribución da demanda dos clientes. Ademais, os avances tecnolóxicos permiten coñecer os niveis de inventario dos socios da cooperativa. Con esta información, a empresa pode seleccionar os clientes aos que atender cada día en función da frota dispoñible. Con todo, necesitan unha ferramenta que se encargue de planificar as rutas dos vehículos para que xeren o mínimo custo de transporte. Para abordar este problema de forma xeral, o noso modelo considerará unha frota de vehículos heteroxénea, tanto en capacidade como en número e capacidade dos seus compartimentos.

## 4. CONCLUSIÓNS

Neste traballo, introducimos unha clase xeral de problemas de rutas de vehículos, nos que a frota é heteroxénea con respecto á capacidade e os vehículos están compartimentados. A frota encárgase de distribuír varios produtos entre un conxunto de clientes, as demandas son estocásticas, cada compartimento non pode conter diferentes produtos ou produtos para diferentes clientes, algúns vehículos non poden acceder a certos clientes e, finalmente, os pedidos dalgúns clientes considéranse urxentes. Propoñemos un algoritmo tabú de optimización, que parte dunha solución obtida mediante un procedemento construtivo. Na nosa opinión, a principal achega deste traballo é a proposta dun modelo estocástico de programación lineal entera mixta, que representa situacións que aparecen na vida real, considerando, de forma nova, diversas restricións que se producen. Xeneralízase o algoritmo de Clarke e Wright e combínase cunha búsqueda tabú para rutas con demandas estocásticas e vehículos con compartimentos, nos que a literatura existente é escasa. Ademais, ilustramos os algoritmos con pequenos exemplos e aplicamos o seu funcionamento a un caso real. Finalmente, deseñamos unha colección de casos de proba para o seu uso en futuras investigacións neste campo. Estes algoritmos están programados no linguaxe *R*, polo que poden integrarse nunha ferramenta máis completa de apoio á decisión dos xestores dunha empresa, xunto con outras, por exemplo, as que prevén o consumo nas explotacións ou xeran informes de interese. Tamén é viable a creación de interfaces sinxelas que faciliten a tarefa dos xestores ou permitan realizar simulacións a usuarios avanzados, mediante o uso de determinadas librerías de *R*. Ditas ferramentas supoñen unha vantaxe para as empresas, xa que permiten a automatización de certas tarefas, por encima doutras metodoloxías "manuais", que son menos eficientes, ou de certas ferramentas comerciais, que son menos transparentes, caras e difíciles de adaptar. Este traballo foi motivado polas tarefas desenvolvidas nunha fábrica de pensos, situada nunha rexión española duns 13.000 $km^2$ de extensión, na que existen máis de 15 empresas deste tipo. O modelo é aplicable a outro tipo de empresas, como as que recollen leite, ou distribúen cereais ou gasolina, e estes algoritmos poden

xerar aforros no transporte que realizan estas empresas. En canto a posibles liñas de investigación futuras, pode ser interesante adaptar o modelo presentado a outros tipos de meta-heurísticas coñecidas, como o recocido simulado ou os algoritmos xenéticos, para logo realizar comparacións cos algoritmos deste traballo utilizando os nosos *benchmarks*. En canto aos modelos alternativos, merece a pena considerar o chamado "problema do camión e o remolque" (véxase Derigs et al., 2013, ou Davila-Pena et al., 2021) para tratar as restricións de accesibilidade, cando devanditos vehículos están ao dispor das empresas e poden reducir os custos de transporte.

## AGRADECEMENTOS

Este traballo foi financiado por European Regional Development Fund (MINECO/AEI grants MTM2017-87197-C3-1-P and C3-3-P).

## REFERENCIAS

Ambrosino, D. and Sciomachen, A. (2007) A food distribution network problem: a case study. IMA Journal of Management Mathematics, 18, 33-53.
Archetti, C., Campbell, A. M. and Speranza, M. G. (2016) Multicommodity vs. single-commodity routing. Transportation Science, 50, 2, 461-472.
Avella, P., Boccia, M. and Sforza, A. (2004) Solving a fuel delivery problem by heuristic and exact approaches. European Journal of Operational Research, 152, 1, 170-179.
Berhan, E., Beshah, B., Kitaw, D. and Abraham, A. (2014) Stochastic vehicle routing problem: A literature survey. Journal of Information & Knowledge Management, 13, 03, 1450022.
Bertsimas, D. J. and Simchi-Levi, D. (1996) A new generation of vehicle routing research: robust algorithms, addressing uncertainty. Operations Research, 44, 286-303.
Brown, G. G., Ellis, C. and Graves, G. W. (1987) Real-time, wide area dispatch of mobil tank trucks. Interfaces, 17, 1, 107-120.
Brown, G. G. e Graves, G. W. (1981) Real-time dispatch of petroleum tank trucks. Management Science, 27, 1, 19-32.
Caramia, M. and Guerriero, F. (2010) A milk collection problem with incompatibility constraints. Interfaces, 40, 2, 130-143.
Chajakis, E. D. and Guignard, M. (2003) Scheduling deliveries in vehicles with multiple compartments. Journal of Global Optimization, 26, 1, 43-78.
Clarke, G. and Wright, J. (1964) Scheduling of vehicles from a central depot to a number of delivery points. Operations Research, 12, 4, 568-581.
Coelho, L. C. and Laporte, G. (2015) Classification, models and exact algorithms for multi-compartment delivery problems. European Journal of Operational Research, 242, 3, 854-864.
Cordeau, J. F., Laporte, G., Savelsbergh, M. W. and Vigo, D. (2007) Vehicle routing. In Handbooks in Operations Research and Management Science, Vol. 14, Barnhart, C., Laporte, G., Eds. Elsevier: North-Holland; pp. 367-428.
Dantzig, G. B. and Ramser, J. H. (1959) The truck dispatching problem. Management Science 6, 1, 80-91.
Davila-Pena, L., Rodríguez-Penas, D. and Casas-Méndez, B. (2021). Novel two-phase heuristic for a new problem of feed distribution with compartmentalized trucks and trailers. Documento de traballo, 30 p. Universidade de Santiago de Compostela.
Derigs, U., Pullmann, M. and Vogel, U. (2013) Truck and trailer routing problems, heuristics and computational experience. Computers & Operations Research, 40, 536-546.
Dror, M., Ball, M. O. and Golden, B. L. (1985) Computational comparison of algorithms for inventory routing. Annals of Operations Research, 4, 3-23.
Dror, M. (2016) Vehicle routing with stochastic demands: models and computational methods. International Series in Operations Research & Management Science, 46, 625-649.
Dror, M. and Trudeau, P. (1986) Stochastic vehicle routing with modified savings algorithm. European Journal of Operational Research, 23, 2, 228-235.
Elbek, M. and Wohlk, S. (2016) A variable neighborhood search for the multi-period collection of recyclable materials. European Journal of Operational Research, 249, 2, 540-550.
El Fallahi, A., Prins, C. and Calvo, R. W. (2008) A memetic algorithm and a tabu search for the multi-compartment vehicle routing problem. Computers & Operations Research, 35, 5, 1725-1741.

Gendreau, M., Laporte, G. and Séguin, R. (1996) Stochastic vehicle routing. European Journal of Operational Research, 88, 1, 3-12.
Glover, F. (1989) Tabu search-part I. ORSA Journal of Computing, 1, 3, 190-206.
Glover, F. (1990) Tabu search-part II. ORSA Journal of Computing, 2, 1, 4-32.
Goodson, J. C. (2015) A priori policy evaluation and cyclic-order-based simulated annealing for the multicompartment vehicle routing problem with stochastic demands. European Journal of Operational Research, 241, 2, 361-369.
Guitián de Frutos, R. M. and Casas-Méndez, B. V. (2019) Routing problems in agricultural cooperatives: A model for optimization of transport vehicle logistics. IMA Journal of Management Mathematics, 30, 387-412.
Henke, T., Speranza, M. G. and Wäscher, G. (2019) A branch-and-cut algorithm for the multi-compartment vehicle routing problem with flexible compartment sizes. Annals of Operations Research, 275, 321-338.
Hsu, C.-I., Hung, S.-F. and Li, H.-C. (2007) Vehicle routing problem with time-windows for perishable food delivery. Journal of Food Engineering, 80, 465-475.
Huang, S. H. (2015) Solving the multi-compartment capacitated location routing problem with pickup-delivery routes and stochastic demands. Computers and Industrial Engineering, 87, 104-113.
Kandiller, L., Eliiyi, D.T. and Tasar, B. (2015) A multi-compartment vehicle routing problem for livestock feed distribution. In Operations Research Proceedings 2015, Selected Papers of the International Conference of the German, Austrian and Swiss Operations Research Societies, University of Vienna, Austria, September 1-4, 2015; Doener, K.F., Ljubic, I., Pflug, G., Tragler, G., Eds. Springer International Publishing, Cham, Switzerland, 149-155.
Kuznietsov, K. A., Gromov, V. A. and Skorohod, V. A. (2016) Cluster-based supply chain logistics: a case study of a Ukrainian food distributor. IMA Journal of Management Mathematics, 28, 553-578.
Laporte, G. (2009) Fifty years of vehicle routing. Transportation Science, 43, 4, 408-416.
Mendoza, J. E., Castanier, B., Guéret, C., Medaglia, A. L. and Velasco, N. (2011) Constructive heuristics for the multicompartment vehicle routing problem with stochastic demands. Transportation Science, 45, 3, 346-363.
Oppen, J. and Lokketangen, A. (2008) A tabu search approach for the livestock collection problem. Computers & Operations Research, 35, 10, 3213-3229.
Ostermeier, M.; Hübner, A. (2018) Vehicle selection for a multi-compartment vehicle routing problem. European Journal of Operational Research, 269, 682-694.
Oyola, J., Arntzen, H. and Woodruff, D. (2017) The stochastic vehicle routing problem, a literature review, part II: solution methods. EURO Journal on Transportation and Logistics, 6, 4, 349-388.
Oyola, J., Arntzen, H. and Woodruff, D. (2018) The stochastic vehicle routing problem, a literature review, part I: models. EURO Journal on Transportation and Logistics, 7, 3, 193-221.
Pandelis, D. G., Kyriakidis, E. G. and Dimitrakos, T. D. (2012) Single vehicle routing problems with a predefined customer sequence, compartmentalized load and stochastic demands. European Journal of Operational Research, 217, 2, 324-332.
Pollaris, H., Braekers, K., Caris, A., Janssens, G. and Limbourg, S. Vehicle routing problems with loading constraints: State-of-the-art and future directions (2015). OR Spectrum, 37, 297-330.
Ruiz, R., Maroto, C. and Alcaraz, J. (2004) A decision support system for a real vehicle routing problem. European Journal of Operational Research, 153, 3, 593-606.
Silvestrin, P. V. and Ritt, M. (2017) An iterated tabu search for the multi-compartment vehicle routing problem. Computers & Operations Research, 81, 192-202 .
Tatarakis, A. and Minis, I. (2009) Stochastic single vehicle routing with a predefined customer sequence and multiple depot returns. European Journal of Operational Research, 197, 2, 557-571.
van der Bruggen, R., Gruson, L. and Salomon, M. (1995) Reconsidering the distribution structure of gasoline products for a large oil company. European Journal of Operational Research, 81, 3, 460-473.
Xia, Y., Fu, Z., Pan, L. and Duan, F. (2018) Tabu search algorithm for the distance-constrained vehicle routing problem with split deliveries by order. Plos One, 13, 5, e0195457. Available online: https://doi.org/10.1371/journal.pone.0195457 (accessed on 22/02/2021).

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# R/EXAMS ARREDOR DO MUNDO: UNHA CONTRIBUCIÓN PARA O SEU USO COA LINGUA GALEGA

Marta Sestelo[1] y Nora M. Villanueva[1]

[1]Departamento de Estatística e I.O., CINBIO & Grupo SiDOR, Universidade de Vigo.

## RESUMO

O paquete `exams` (Zeileis et al., 2020) para a linguaxe de programación e entorno de software libre para computación estatística R proporciona un enfoque único e xeral para a xeración de exames automáticos. Mediante o uso de plantillas de exercicios dinámicos pódense crear gran cantidade de exames ou probas para varios sistemas: PDF para exames escritos, formatos de importación para sistemas de xestión da apredizaxe (como Moodle, Canvas, etc.), votación en directo (a través de ARSnova) ou saídas personalizadas en PDF, HTML, DOCx, etc.

O software R/exams (Gruen and Zeileis, 2009; Zeileis et al., 2014) ofrece un formato estandarizado chamado "NOPS"para exames escritos con exercicios (de opción múltiple e/ou de elección única) que se poden xerar, escanear e avaliar automaticamente. Para que a dixitalización automática funcione correctamente, a páxina de título ten un deseño fixo que os usuarios non poden modificar. Sen embargo, mediante o uso do argumento `language` pódese cambiar o soporte da linguaxe. Neste traballo preséntase a contribución realizada para o uso da lingua galega e móstrase a sua aplicación por medio dalgúns exemplos.

**Palabras e frases chave**: R, exames, docencia, galego, sistemas de xestión da aprendizaxe, e-learning

## REFERENCIAS

Gruen, B. and Zeileis, A. (2009). Automatic Generation of Exams in R. Journal of Statistical Software 29(10), 1–14.

Zeileis, A., Umlauf, N. and Leisch, F. (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. Journal of Statistical Software 58(1), 1–36.

Zeileis, A., Gruen, B., Leisch, F., Umlauf, N., Birbaumer, M., Ernst, D., Keller, P., Smits, N., Stauffer, R., Sato, K. (2020). exams: Automatic Generation of Exams in R. R package version 2.3-6.

# Concurso categoría A

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# BOUND-TIGHTENING

Ignacio Gómez-Casares[1]

[1] Estudiante del Máster en Técnicas Estadísticas

## RESUMEN

El bound-tightening agrupa a un conjunto de técnicas que se aplican en los problemas de programación matemática para reducir el dominio de las variables y así mejorar el funcionamiento de los algoritmos utilizados para resolverlos. En este póster presento las dos técnicas de bound-tightening que están implementadas en el solver RAPOSa junto con algún resultado numérico que pone de manifiesto el impacto que tienen las mismas sobre el tiempo de ejecución.

**Palabras y frases clave**: Bound-tightening; Programación matemática; Investigación operativa; OBBT; FBBT; RAPOSa

## REFERENCIAS

BELOTTI, P, S CAFIERI, J LEE a L LIBERTI, 2012. On feasibility based bounds tightening. hal-enac.archives-ouvertes.fr

BELOTTI, Pietro, Jon LEE, Leo LIBERTI, François MARGOT a Andreas WÄCHTER, 2009. Branching and bounds tightening techniques for non-convex MINLP. Optimization Methods and Software. 24(4-5), 597-634.

BUSSIECK, Michael R., Arne Stolbjerg DRUD a Alexander MEERAUS, 2003. MINLPLib—A Collection of Test Models for Mixed-Integer Nonlinear Programming. INFORMS Journal on Computing. 15(1), 114-119.

DOLAN, Elizabeth D. a Jorge J. MORÉ, 2002. Benchmarking optimization software with performance profiles.Springer Science and Business Media LLC.

GLEIXNER, Ambros M., Timo BERTHOLD, Benjamin MÜLLER a Stefan WELTGE, 2017. Three enhancements for optimization-based bound tightening. Journal of Global Optimization. 67(4), 731-757.

PURANIK, Yash a Nikolaos V. SAHINIDIS, 2017. Domain reduction techniques for global NLP and MINLP optimization. Constraints. 22(3), 338-376.

SHERALI, Hanif D. a Cihan H. TUNCBILEK, 1992. A global optimization algorithm for polynomial programming problems using a Reformulation-Linearization Technique. Journal of Global Optimization. 2(1), 101-112.

# Concurso categoría B

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# DESENVOLVEMENTO DUN INDICADOR DE ALTA FRECUENCIA PARA O SEGUIMENTO DA ECONOMÍA ESPAÑOLA

Lucía Gil Rial

## RESUMO

Dende o Departamento de Planificación Estratéxica e PMO de ABANCA buscan desenvolver un indicador da economía no país baseado en métricas de alta frecuencia. Este indicador permitiría anticipar o comportamento da actividade económica nun prazo de tempo inferior que os indicadores tradicionais, como o Produto Interior Bruto (PIB), que adoitan ter unha frecuencia trimestral e, ademais, existen atrasos na súa publicación.

Neste traballo dáse comezo á construción dun indicador baseado na demanda de electricidade diaria total en España e compárase a súa dinámica coa doutros indicadores usuais, como o PIB, resaltando a relación entre estes indicadores ao longo do 2020. Estes resultados poden verse nunha aplicación elaborada en *Shiny*.

Así, mediante a aplicación de técnicas de series de tempo e números índice, desenvolvemos un indicador de alta frecuencia que proporciona información sobre grandes impactos na economía en tempo real, esquivando a espera doutros indicadores usuais de frecuencias máis baixas.

**Palabras e frases chave:** Indicador macroeconómico, series temporais, números índice, R, *Shiny*.

## 1. INTRODUCIÓN

A pandemia provocada pola COVID-19 puxo de manifesto a necesidade de dispoñer de indicadores de alta frecuencia para analizar a evolución do estado económico ante situacións tan abruptas. Como punto de partida para a construción deste indicador, tomamos a demanda eléctrica diaria total en España. Previo ao desenvolvemento do mesmo, realizamos unha análise exploratoria da demanda eléctrica ao longo dos anos. Unha vez que se estudou a dinámica da serie, limpámola de observacións atípicas e dos efectos semanal e mensual, este último provocado polo cambio de temperaturas. A continuación, construímos un indicador baseado na demanda eléctrica limpa e comparamos a súa dinámica coa doutros indicadores macroeconómicos tradicionais. Por último, deseñamos unha aplicación en *Shiny*, na que se mostran os resultados deste indicador.

## 2. ANÁLISE DA DEMANDA ELÉCTRICA

O noso obxectivo é a creación dun indicador de alta frecuencia para avaliar o estado económico en España. Para isto, nunha primeira instancia, consideramos a demanda eléctrica diaria total no país. Resulta de gran interese coñecer a dinámica desta serie temporal ao longo dos anos, polo que facemos tres estudos:

1. Analizamos o ano 2017, que tomamos como mostra dun ano usual en termos de evolución económica.
2. Analizamos a dinámica da serie a longo prazo, considerando o período temporal entre o 2015 e o 2019.
3. Analizamos o ano 2020, cuxo comportamento resulta ser atípico, como consecuencia do confinamento.

No primeiro caso observamos que a demanda eléctrica presenta un efecto semanal, constituído por unha maior demanda eléctrica nos días entre semana, un efecto mensual, onde a temperaturas máis extremas hai unha maior demanda de electricidade, e observacións atípicas en festivos, como o Nadal ou Semana Santa.

No segundo estudo podemos ver que o nivel da demanda eléctrica non cambia substancialmente nestes anos e que a súa dinámica anualmente coincide coa xa comentada.

Por último, esta evolución anual da demanda eléctrica rómpese no 2020, debido á pandemia de COVID-19. O confinamento provocou que todas as actividades non esenciais se interrompesen, polo que a demanda eléctrica sufriu un gran descenso neste período, tal e como se pode ver na Figura 1. Deste feito podemos concluír que a demanda de electricidade capta os cambios bruscos no estado da economía e, así, a súa selección para a construción do indicador parece ser axeitada.

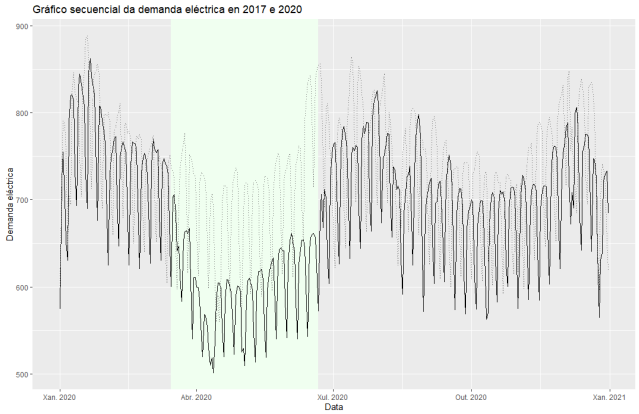



Figura 1: Gráficos secuenciais da demanda eléctrica en España en 2020 en negro e en 2017 en liña punteada. A zona sombreada correspóndese co confinamento.

## 3. LIMPEZA DA DEMANDA ELÉCTRICA

Para a construción do indicador, interésanos traballar só con aquela información da demanda eléctrica relacionada coa actividade económica e, por iso, debemos corrixir as observacións atípicas e os efectos semanal e mensual.

1. Substituímos os atípicos que identifica a función *forecast::tsoutliers()* polas súas propostas e, a maiores, dado que esta función non capta o cambio de nivel na semana de Nadal, substituímos estes datos polos da semana anterior, por sinxeleza e para manter a dinámica semanal.
2. O efecto semanal pode corrixirse mediante medias móbiles con xanela 7.
3. Para a corrección do efecto mensual probamos tres metodoloxías:
    a) O axuste dun modelo de regresión que explicase a relación entre a demanda eléctrica e a temperatura.
    b) O axuste de varios modelos de regresión que explicasen a relación entre a variación da demanda eléctrica e a temperatura.
    c) A variación mensual respecto da mediana da demanda eléctrica.

    As dúas primeiras metodoloxías non deron uns resultados adecuados, mentres que a terceira proposta consegue corrixir o efecto mensual.

## 4. CONSTRUCIÓN DO INDICADOR

Unha vez que dispoñemos da demanda eléctrica limpa, construímos un indicador en base 100 tomando como período base o 2015 e de frecuencias diaria, semanal e mensual. Ademais, comparamos a súa dinámica co Produto Interior Bruto, o Índice de Produción Industrial e o Índice da Rede Eléctrica, tendo en conta que o noso indicador só se basea nunha única variable e, polo tanto, non é esperable que a súa dinámica coincida coa destes agregados. Aínda así, podemos ver que existe certa relación entre eles ao longo do 2020, tal e como se mostra na Figura 2.

O noso indicador capta o grande impacto que provocou a pandemia na economía e danos unha medida deste impacto en tempo real, adiantándose ao PIB, cuxa frecuencia é trimestral e, ademais, ten un desfase de publicación de aproximadamente 50 días.

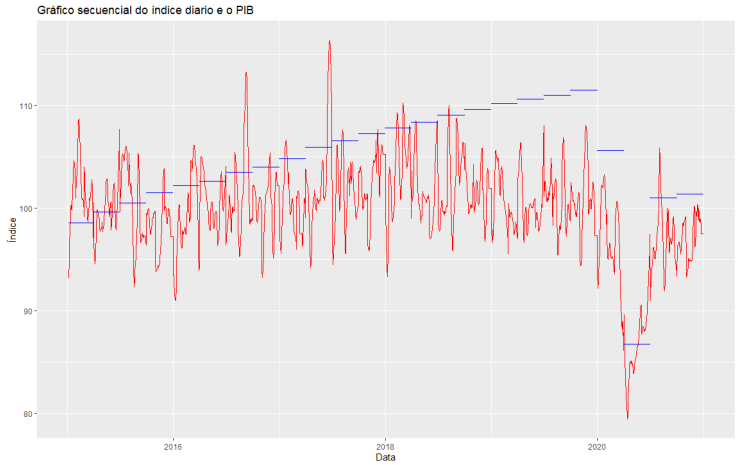



Figura 2: Gráfico secuencial do indicador diario en vermello e do PIB en azul.

## 5. APLICACIÓN EN SHINY

Desenvolvemos unha aplicación en *Shiny* para presentar os resultados do traballo a todo usuario de ABANCA non coñecedor de R.

Unha vez que se executa a aplicación, descárganse automaticamente os datos de demanda eléctrica diaria en España mediante a API da Rede Eléctrica e móstranse os gráficos secuencias da demanda eléctrica e a súa corrección, así como o indicador proposto e a súa comparación coa dos agregados macroeconómicos xa mencionados. Ademais, permítese descargar os datos do indicador segundo a súa frecuencia nun ficheiro *Excel*.

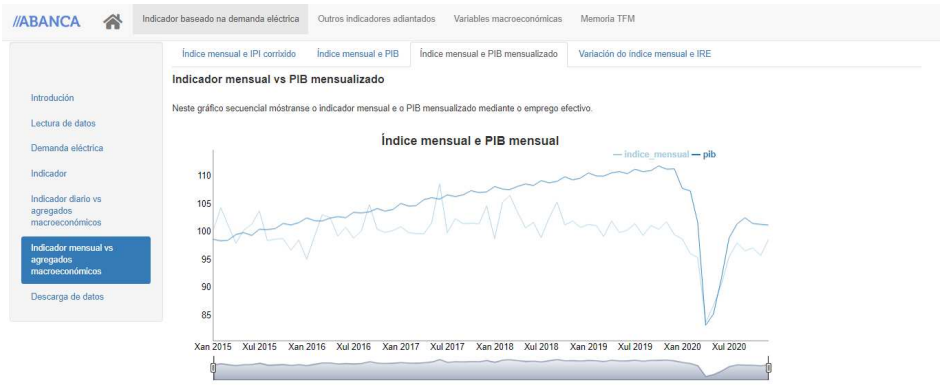



Figura 3: Captura dunha xanela da aplicación.

## 6. CONCLUSIÓNS

Neste traballo damos comezo ao desenvolvemento dun indicador de alta frecuencia para o seguimento da economía española, considerando como base a demanda eléctrica diaria total en España.

Primeiro, estudamos a dinámica da demanda eléctrica, concluíndo que reflicte os cambios bruscos na economía, e corrixímola de atípicos e dos efectos semanal e mensual.

Logo, a partir da demanda eléctrica limpa, construímos un indicador de diferentes frecuencias e comparamos a súa evolución coa doutros indicadores máis tradicionais, onde destaca o seu parecido no 2020. En consecuencia, acadamos un indicador de alta frecuencia que permite identificar grandes cambios na actividade económica en tempo real e, amais, anticipa o aviso desta anomalía respecto aos indicadores usuais.

Por último, fixemos unha aplicación en *Shiny*, para que todo usuario en ABANCA poida coñecer os resultados do indicador.

Neste proxecto dáse un primeiro paso na solución do problema que propón ABANCA, que se enmarca no mundo multivariante. Polo tanto, o seguinte paso, no que xa estamos traballando, reside en engadir máis variables de alta frecuencia, como a mobilidade, os pagos con tarxetas e as tendencias extraídas de Google, para chegar a un indicador que inclúa unha información máis completa da actividade económica e que permita anticipar os seus cambios.

## REFERENCIAS

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2020) shiny: Web Application Framework for R. R package version 1.4.0.2. URL: *https://CRAN.R-project.org/package=shiny*.

Cryer JD, Chan K (2010) Time Series Analysis. With Applications in R. Springer, Nova York.

Peña D, Romo J (1997) Introducción a la Estadística para las Ciencias Sociales. McGraw-Hill, Madrid.

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

*https://www.ree.es/es*.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# OPTIMIZACIÓN DO CIRCUITO DE PACIENTES DO HOSPITAL DE DÍA DE ONCOLOXÍA

Adrián González Maestro[1]

[1] FIDIS, Hospital Clínico Universitario de Santiago de Compostela

**RESUMO**

O obxectivo deste traballo é usar a información da que dispoñemos dos pacientes do hospital de día de oncoloxía de Santiago para distribuir as distintas citas médicas que teñen en cada unha das súas visitas a dito centro de xeito que minimicemos a duración das estancias na sala de espera do conxunto total dos pacientes e tendo en conta as restriccións do circuito.

**Palabras e frases chave**: Programación matemática, quimioterapia ambulatoria.

## 1. INTRODUCIÓN

O hospital de día de oncoloxía de Santiago de Compostela encárgase de ofrecer servizos médicos ambulatorios aos pacientes de cancro que non están hospitalizados. Durante as súas visitas a estes centros, a gran maioría dos pacientes reciben dous servizos médicos: unha revisión oncolóxica e un tratamento de quimioterapia. Ditos servizos están fortemente relacionados, pois a revisión oncolóxica é esencialmente un requisito previo no que compróbase se a saúde do paciente é apta para recibir o tratamento de quimioterapia, que é o obxectivo final da visita ao centro. Entre a realización de ambos servizos médicos, o paciente debe realizar unha estancia na sala de espera do hospital, séndolle imposible abandonar o centro durante ese tempo pois descoñece en que momento será chamado para comezar o seu tratamento de quimioterapia.

De cada paciente coñécese, a priori, o horario no que en teoría vai comezar a súa revisión oncolóxica e canto vai durar o seu tratamento de quimioterapia unha vez este comece. O problema abordado neste traballo é idear un algoritmo que, facendo uso desa información e considerando un conxunto de pacientes dunha xornada laboral concreta, ofreza un horario estimado para o comezo da quimioterapia de cada un dos pacientes de xeito que se minimicen as esperas dos mesmos entre ditos compromisos. Deste xeito estariamos conseguindo dous propósitos, por un lado minimizar os tempos mortos dos pacientes entre ambos compromisos médicos, e polo outro darlle ao paciente a tranquilidade de ter unha estimación do horario do comezo do seu tratamento de quimioterapia, a cal lle permitiría incluso abandonar o hospital durante dito período horario.

## 2. DESCRIPCIÓN DO PROBLEMA

Describiremos con detalle o circuito que experimentan a gran maioría dos pacientes nas súas visitas a este centro. En primeiro lugar rexístrase a súa chegada e realízaselle unha análise de sangue. Cando os resultados están listos, un oncólogo certifica que estes son correctos e ten unha revisión co paciente. Se todo o relacionado coa saude do paciente está en orde, ordénase á farmacia a preparación das substancias da quimioterapia do paciente. Cando ditas substancias están listas, lévase ao paciente a un dos sillóns de infusión (no caso de que haxa un libre) e realízase a infusión das substancias. Cando o proceso remata o paciente deixa o sillón e abandona o centro. Representamos este proceso na Figura 1.

Durante todo este proceso existen pasos que non involucran ao paciente, e durante os cales este debe agardar ata o comezo da seguinte etapa. Referímonos ao proceso de estudo da analítica e á

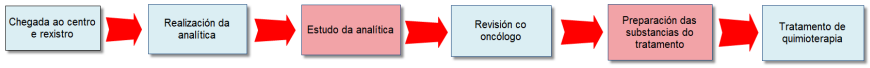



Figura 1: Circuito seguido polos pacientes oncolóxicos

preparación das substancias que serán suministradas ao paciente. Como xa mencionamos anteriormente, a segunda destas etapas de agarda presenta o inconveniente de que o paciente non coñece unha estimación de en que momento comezará o seu tratamento de quimioterapia, o cal dota a este período de espera dunha especial incerteza. Ademáis, os tempos de agarda da segunda etapa son sensiblemente maiores que os da primeira. Concretamente, a primeira espera dura normalmente entre 45 e 60 minutos, mentres que a segunda ten un promedio de duración de 110 minutos. Este é o motivo polo cal o algoritmo que deseñemos neste traballo terá por obxectivo organizar os horarios das sesións de quimioterapia.

Agora, antes de definir formalmente un algoritmo que organice os horarios dos pacientes, debemos ter en conta todas as restriccións que debería respetar a solución que dito algoritmo nos ofreza. No noso caso existen restriccións horarias e de capacidade. En primeiro lugar, con respecto ás restriccións horarias temos que o centro abre ás 8:00 h e pecha ás 22:00 h, o cal implica que ningún tratamento pode comezar antes das oito da mañá nin rematar máis tarde das dez da noite.

Con respecto ás restriccións de capacidade temos que ter en conta que a sala de quimioterapia do hospital conta con 40 sillóns para levar a cabo tratamentos de quimioterapia, o cal implica que non é posible que teñan lugar máis de 40 tratamentos simultaneamente. Tamén existen unha serie de restriccións deste tipo relacionadas coa capacidade de traballo dos enfermeiros. Estes profesionais son os encargados de inicializar os tratamentos dos pacientes e de vixialos durante o seu transcurso. Cada enfermeiro pode encargarse da posta en marcha dun tratamento cada 15 minutos (que é o tempo necesario para desinfectar o sillón, ubicar ao paciente no mesmo, poñerlle a vía, etc) e, simultaneamente, pode vixiar ata un máximo de 16 tratamentos en curso. Esas son as restriccións de capacidade aplicadas a estes profesionais na sala de quimioterapia, sendo o número de enfermeiros dispoñibles un parámetro que evoluciona ao longo da xornada da seguinte forma: están operativos 5 enfermeiros de 8:00 h a 10:00 h, 6 de 10:00 h a 15:00 h, 3 de 15:00 h a 17:00 h e 2 de 17:00 h a 22:00 h.

## 3. REVISIÓN DA LITERATURA

Existen multitude de traballos abordando labores de optimización e programación en circunstancias similares ás que acabamos de presentar. Algúns exemplos son Heshmat e Eltawil (2019), Turkcan et al. (2012), Liang et al. (2015) ou Hesaraki et al. (2019).
Eses dous últimos estudos foron os que inspiraron en maior medida o algoritmo que a continuación presentaremos. En Liang et al resólvese o problema de planificar as citas dos pacientes que acuden a un centro de oncoloxía tendo como principal obxectivo acadar un esquema horario que evite picos na carga de traballo dos profesionais do centro durante a xornada laboral, mentres que en Hesaraki et al persíguese principalmente reducir as esperas dos pacientes dentro da medida do posible. No que á función obxectivo se refire, o noso algoritmo aseméllase máis ao traballo de Hesaraki et al, se ben o conxunto de restriccións xeradas polo contexto dese estudo difire bastante das que emanan da realidade do hospital de Santiago.

## 4. IMPLANTACIÓN DO ALGORITMO

Xa coñecemos as restriccións que a solución ao problema debe cumprir, polo tanto procedemos a describir a idea que persigue o algoritmo no seu funcionamento. Posteriormente definirémolo formalmente. Como xa mencionamos anteriormente, de cada paciente coñecemos o momento no que teoricamente comezaría a súa revisión oncolóxica e canto durará o seu tratamento de quimioterapia. O que descoñecemos é canto se retrasará dita revisión oncolóxica, canto durará e canto tempo transcurrirá durante a preparación das substancias para a quimioterapia do paciente. Se

puidésemos saber o valor deses tres tempos para cada paciente, saberíamos a partir de que momento cada paciente estaría listo para recibir o seu tratamento de quimioterapia. Como iso non é posible, o que fixemos foi recoller datos no hospital para esas tres variables sobre o conxunto de pacientes dunha semana de traballo no centro. A idea é analizar os datos recollidos e establecer unha cota superior da suma das tres variables mencionadas para unha porcentaxe razoablemente alta dos pacientes. Deste xeito, se tomamos as horas de inicio teóricas das revisións oncolóxicas dos pacientes e sumámoslles esta cota, obteremos unhas horas a partir das cales deberían estar listos para recibir os seus tratamentos de quimioterapia (nalgún caso pode que non sexa así pero, novamente, a idea é que funcione nun número razoablemente alto de pacientes). Tras analizar os datos recollidos no hospital, decidimos traballar con catro posibles cotas: 210 minutos, 180 minutos, 150 minutos e 120 minutos. A primeira delas é unha cota, en xeral, demasiado conservadora, que sería representativa de días de traballo nos que o fluxo de pacientes fose máis lento do habitual. A segunda é a cota máis realista e adecuada para xornadas de traballo usuais. Finalmente, as cotas de 150 e 120 minutos presupoñen unha velocidade do circuito de pacientes superior ao usual. Serán utiles para comprobar o que poderiamos acadar usando o algoritmo que imos propoñer en escenarios onde a fluidez dos pacientes polas distintas etapas do proceso fose maior do que a día de hoxe é normal. É dicir, serán útiles para comprobar cuan mellores serían as programacións ofrecidas polo algoritmo se o hospital conseguise, dalgunha forma, reducir en 30 ou 60 minutos a suma media dos tempos do retraso da cita co oncólogo, a duración de dita cita e a preparación das substancias para a quimioterapia.

Finalmente, antes de plantexar o modelo é preciso mencionar que, para poder modelar matematicamente a situación precisamos discretizar as 14 horas de traballo dunha xornada laboral deste centro nunha certa cantidade de intervalos finitos de tempo. Facer isto representa unha perda de precisión á hora de traballar co parámetro tempo, e canto máis grandes sexan eses intervalos peor será a solución obtida. Non obstante, se ditos intervalos son excesivamente pequenos, non serían prácticos de cara a crear un esquema organizativo con eles, pois presuporían unha puntualidade das distintas etapas do proceso que en xeral non se cumpriría. Decidimos que unha amplitude de 5 minutos é un bo compromiso entre precisión e manexabilidade dos intervalos. Polo tanto dividimos as 14 horas da xornada laboral en 168 intervalos de 5 minutos. Agora que temos unha estimación para cada paciente de a partires de que momento estará listo para recibir o seu tratamento de quimioterapia e dispoñemos dunha cantidade discreta de intervalos horarios nos cales programar ditos tratamentos, estamos en condicións de levar a cabo o proceso de minimización. Para iso plantexamos o seguinte problema de programación matemática.

## 5. MODELADO DO PROBLEMA

En primeiro lugar describimos o conxunto de parámetros e variables:

# Parámetros:

$P$: parámetro enteiro que determina o número de pacientes que teñen asignado un tratamento de quimioterapia durante a xornada de traballo.
$l_p$ con $p \in \{1, ..., P\}$: duración do tratamento do paciente $p$.
$r_p$ con $p \in \{1, ..., P\}$: intervalo de tempo a partires do cal o tratamento do paciente $p$ pode comezar.
$N$: número de enfermeiros traballando durante a xornada.
$Ndisp_t$ con $t \in \{1, ..., T\}$: número de enfermeiros dispoñibles durante o intervalo horario $t$.

# Variables:

$x_{it}$ con $i \in \{1, ..., P\}$ e $t \in \{1, ..., 168\}$: variable binaria que toma o valor 1 se o paciente $i$ comeza o seu tratamento no período horario $t$.
$C_{max}$: variable enteira que determina a partires de que franxa horaria non se levarán a cabo máis tratamentos.
$\lambda_1$ e $\lambda_2$: dous números reais que usaremos para ponderar na nosa función obxectivo. Cumpren que

$\lambda_1 + \lambda_2 = 1$ e $\lambda_1, \lambda_2 \in [0,1]$.

Procedemos finalmente a definir formalmente o modelo:

minimizar: $\lambda_1 \cdot C_{max} + \lambda_2 \cdot \sum_{p=1}^{P} \sum_{t=1}^{168} [(t - 1 - r_p) \cdot x_{p,t}]$

suxeito a:

$$\sum_{t=1}^{168} x_{p,t} = 1, \qquad \forall p \in \{1, ..., P\} \quad (1)$$

$$\sum_{t=1}^{r_p} x_{p,t} = 0, \qquad \forall p \in \{1, ..., P\} \text{ tal que } r_p > 0 \quad (2)$$

$$C_{max} \leq 168 \quad (3)$$

$$\sum_{t=1}^{168} (t + l_p - 1) \cdot x_{p,t} \leq C_{max}, \qquad \forall p \in \{1, ..., P\} \quad (4)$$

$$\sum_{p=1}^{P} \sum_{a=max\{1,t-l_p+1\}}^{t} x_{p,a} \leq 40, \qquad \forall t \in \{1, ..., 168\} \quad (5)$$

$$\sum_{p=1}^{P} x_{p,t} \leq Ndisp_t, \qquad \forall t \in \{1, ..., 168\} \quad (6)$$

$$\sum_{p=1}^{P} [\sum_{a=max\{1,t-l_p+1\}}^{t} x_{p,a}] \leq 16 \cdot Ndisp_t, \qquad \forall t \in \{1, ..., 168\} \quad (7)$$

Como vemos, a función obxectivo é o sumatorio dunha expresión conformada por dous sumandos ponderados. O primeiro deles fai referencia ao momento horario no que remata o derradeiro tratamento. O segundo termo é unha expresión que crece conforme aumenta o tempo transcurrido entre o momento no que o tratamento do paciente puido comezar e o momento no que realmente comezou. Minimizar o primeiro destos sumandos significa escoller un esquema horario que remate a derradeira sesión de quimioterapia o antes posible, mentres que minimizar o segundo implica diminuir a espera total dos pacientes trala revisión co oncólogo. Ambos obxectivos son desexables, un de cara a mellorar a calidade do servizo para os pacientes e o outro de cara a darlle aos profesionais do centro unha marxe para lidiar cos retrasos que necesariamente danse habitualmente no ámbito sanitario, previndo así a realización de horas de traballo máis alá da hora de peche oficial do centro. A filosofía do hospital é priorizar o benestar dos pacientes, e polo tanto tipicamente tomarase $\lambda_2 >> \lambda_1$ á hora de executar o modelo.

Comentamos agora o conxunto de restriccións. Vemos que a (1) ten por obxectivo asegurar que a todos os pacientes se lles asigna un horario para a sesión de quimioterapia. A (2) impide que un tratamento sexa programado antes de que o paciente estea listo para recibilo. A (3) verifica que a variable $C_{max}$ toma un valor cun horario asociado que non supera as 22:00 h. A (4) impón que todos os tratamentos rematen como moi tarde na hora marcada polo valor que tome a variable $C_{max}$. A (5) impide que nalgún momento haxa máis de 40 tratamentos en curso. A (6) e a (7) están relacionadas ca capacidade de traballo dos enfermeiros. A primeira delas asegura que en ningún momento da xornada inicialízase un número de tratamentos superior ao número de enfermeiros dispoñibles nese momento. A segunda delas impide que en ningún momento algún enfermeiro teña que supervisar máis de 16 tratamentos simultaneamente.

## 6. RESULTADOS

A continuación amosamos a Táboa 1. Nela visualizamos os minutos de agarda experimentados polos pacientes co procedemento actual durante a semana na que tomamos datos no hospital e os minutos de agarda que se terían dado aplicando o algoritmo coas cotas anteriormente mencionadas.

| | Actualmente | Modelo1 | Modelo2 | Modelo3 | Modelo4 | Pacientes |
|---|---|---|---|---|---|---|
| **Luns** | 7.339 | 9.255 | 7.900 | 6.250 | 4.600 | 56 |
| **Martes** | 10.382 | 13.375 | 11.675 | 9.275 | 7.115 | 72 |
| **Mércores** | 7.351 | 9.790 | 8.190 | 6.360 | 4.480 | 61 |
| **Xoves** | 5.906 | 7.790 | 6.360 | 4.860 | 3.360 | 50 |
| **Venres** | 5.810 | 8.075 | 6.615 | 5.085 | 3.555 | 51 |
| **Totais** | 36.788 | 48.285 | 40.740 | 31.830 | 23.110 | 290 |

Táboa 1: Minutos de espera dos pacientes co procedemento actual e cos distintos escenarios do modelo

Na primeira columna figuran os tempos de espera do procedemento actual, mentres que nas seguintes recollemos os tempos de agarda resultantes de aplicar o modelo considerando unha cota concreta. Na columna "Modelo 1" consideramos a cota de 210 minutos. A columna "Modelo 2" representa os datos análogos tomando unha marxe de 180 minutos. Na columna "Modelo 3" tómase unha marxe de 150 minutos e, finalmente, na columna "Modelo 4" tómanse 120 minutos de marxe.

Agora observemos como, na Figura 2, apréciase un grafo onde comparamos a evolución durante a xornada laboral do número de tratamentos en curso na sala de quimioterapia co procedemento actual e co algoritmo usando a cota de 120 minutos.

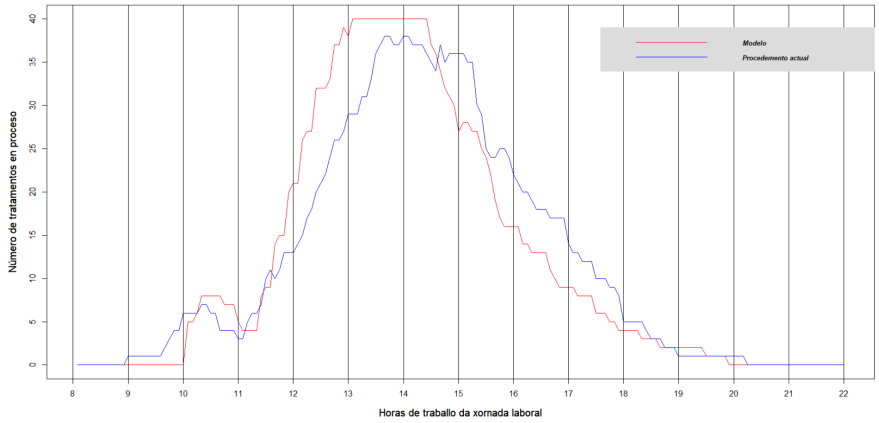



Figura 2: Evolución da carga de traballo dos enfermeiros co procedemento actual e co algoritmo usando a cota de 120 minutos

Como se pode percibir, o noso algoritmo tende a programar máis tratamentos nas horas do mediodía e a aliviar a carga de traballo das horas da tarde. Este feito representa unha ventaxa do uso do algoritmo, dado que ofrece aos profesionais unha maior marxe para lidiar con posibles incidencias que poidan retrasar o transcurso normal do fluxo de pacientes. Isto pode traducirse en que os profesionais do centro non teñan que traballar máis alá das 22:00 h (hora oficial de peche) aínda padecendo retrasos inesperados.

## 7. CONCLUSIÓNS

Se analizamos en detalle a Táboa 1, podemos comprobar que o resultado de aplicar o noso modelo coas cotas de 210 e 180 minutos é aumentar os tempos de agarda dos pacientes a cambio de ofrecerlles de antemán un horario concreto tanto para a súa revisión oncolóxica como para a súa sesión de quimioterapia. Debemos ter en conta que este aumento, no caso da cota de 180 minutos (a máis realista), é de menos de 4000 minutos sobre o conxunto de todos os pacientes da semana

na que se recolleron os datos. Isto representa un aumento na espera en promedio de menos dun cuarto de hora por paciente, o cal non parece demasiado se temos en conta que a espera media por paciente co procedemento actual está próxima ás dúas horas de duración (110 minutos). Polo tanto podería parecer que pagaría a pena aplicar o algoritmo e asumir este incremento na espera para obter uns horarios completamente definidos de antemán. Ademáis, tamén temos que ter en conta que a distribución dos tratamentos obtida co algoritmo ofrece aos profesionais unha maior capacidade de reacción frente a improvistos, como mencionamos ao analizar a Figura 2. Por outra parte, tamén parecería interesante explorar a opción de reducir os tempos da segunda etapa de agarda dos pacientes mediante unha inversión razoable de recursos. Como podemos consultar na Táboa 1, só con reducir en 30 minutos a espera media xa poderíamos obter un esquema organizativo completo para os pacientes sen aumentar os tempos de espera respecto da situación actual. Ao contrario, os tempos veríanse reducidos respecto de dita referenza.

## REFERENCIAS

A. F. Hesaraki, N. P. Dellaert, and T. de Kok (2019). Generating outpatient chemotherapy appointment templates with balanced flowtime and makespan. European Journal of Operational Research 275: 304-318.

M. Heshmat and A. Eltawil (2019). Solving operational problems in outpatient chemotherapy clinics using mathematical programming and simulation. Annals of Operations Research 298: 289-306.

B. Liang, A. Turkcan, M. E. Ceyhan and K. Stuart (2015). Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. International Journal of Production Research Vol 53, No 24: 7177-7190.

A. Turkcan, B. Zeng and M. Lawley (2012). Chemotherapy operations planning and scheduling. IIE Transactions on Healthcare Systems Engineering 2 (1): 31-49.

# Mesas redondas

## O impacto da COVID-19 dende diferentes perspectivas

**Moderadora:**

**Rosa M. Crujeiras Casais (Universidade de Santiago de Compostela)**

**Participantes:**

**Alberto Ruano Raviña (Universidade de Santiago de Compostela)**

**Ricardo Cao Abad (Universidade de A Coruña)**

**José Antonio Campo Andión (Instituto Galego de Estatística)**

**Diego Ramiro Fariñas (Consejo Superior de Investigaciones Científicas)**

A Organización Mundial da Saúde (OMS), o 11 de marzo de 2020, declarou pandemia internacional á situación de emerxencia de saúde pública provocada polo COVID-19, enfermidade causada polo coronavirus SARS- CoV-2, que xurdiu en China a finais de decembro do ano previo. Unha gran cantidade de países sufriron consecuencias inmediatas debido á rápida propagación da enfermidade, entre eles España, que decretou o 14 de marzo de 2020 o estado de alarma. A este feito, sucedéronlle unha serie de medidas restritivas co obxectivo de refrear o avance da enfermidade, como o confinamento domiciliario. Tendo en conta esta situación, o obxectivo desta mesa redonda é avaliar o impacto que supuxo na sociedade o avance da COVID-19 desde diferentes perspectivas como a epidemioloxía, a estatística, a demografía ou a economía. Para isto na mesa participarán expertos que traballaron activamente na avaliación do impacto da COVID-19 nas súas áreas de coñecemento.

## Estatística e IO noutros mundos

**Moderadora:**

**Milagros Diéguez Taboada (IES de Milladoiro, Xunta de Galicia)**

**Participantes:**

**Carlos Amiama Ares (Universidade de Santiago de Compostela)**

**Ángel Carracedo Álvarez (Universidade de Santiago de Compostela)**

**Jesús Lagos Milla (ORANGE)**

**José Ramón Pichel Campos (IMAXIN)**

Ninguén dubida do rol fundamental da estatística na sociedade actual. A gran profusión de datos que se xeran na era das tecnoloxías da información e da comunicación, fan imprescindible o coñecemento e o manexo desta ciencia. Ademais de pola súa vertente intelectual, a estatística goza dun gran recoñecemento como ciencia aplicada, e é ben coñecido que se emprega en diversos campos tales como a economía, nas ciencias sociais ou mesmo na física cuántica. Porén ten outros ámbitos de aplicación menos coñecidos aos que nos queremos achegar dende esta mesa redonda. Nela trataremos de expor como está presente, no mundo dos deportes, no campo da xenética, ou mesmo na filoloxía sen olvidar a Investigación de Operacións aplicada a mellora da loxística no ámbito agrario.

# Obradoiros

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# FERRAMENTAS EN R PARA A RESOLUCIÓN DE PROBLEMAS DE INVESTIGACIÓN DE OPERACIÓNS

Alejandro Saavedra-Nieves[1]

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

## RESUMO

Aínda que moitos dos problemas existentes no eido da Investigación de Operacións se modelen de forma natural e sinxela, é posible que a obtención das súas solucións con lapis e papel non sexa sempre doada. Entre outros, podemos destacar o caso da resolución de problemas de programación lineal, do cómputo de solucións específicas da teoría dos xogos cooperativos, da análise de problemas de inventario, ou da xestión de proxectos. Nestes contextos, o uso de ferramentas informáticas pode ser realmente útil. Neste taller abordarase o manexo dalgunhas das existentes no software R para a resolución de problemas coma os mencionados

**Palabras e frases chave**: Investigación de Operacións, software, R.

## REFERENCIAS

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# SENTIDO ESTADÍSTICO Y SU DESARROLLO MEDIANTE PROYECTOS E INVESTIGACIONES

Carmen Batanero[1] y José A. Garzón-Guerrero [2]

[1] Departamento de Didáctica de la Matemática, Universidad de Granada
[2] Departamento de Didáctica de la Matemática, Universidad de Granada

**RESUMEN**

Hoy en día el ciudadano se enfrenta a una gran variedad de información estadística sobre temas de su interés y con frecuencia ligada a la toma de decisiones. Es importante que la escuela facilite el desarrollo del sentido estadístico, entendido como la unión de la cultura y el razonamiento estadístico, junto con unas actitudes adecuadas. El objetivo de la presentación realizada en el congreso ha sido analizar las componentes del sentido estadístico y mostrar ejemplos de proyectos e investigaciones que permiten desarrollarlo.

**Palabras y frases clave:** sentido estadístico, proyectos, gráficos, educación secundaria.

## 1. INTRODUCCIÓN

La difusión de las tecnologías de la comunicación ha supuesto un cambio drástico en el comportamiento a la hora de consumir información estadística proporcionada por organismos nacionales e internacionales y disponible para su análisis por parte del ciudadano (Ridgway, 2016). Además, gran parte de esta información estadística se presenta en forma gráfica, con la intención de transmitirla de manera rápida y eficiente, a la vez que auto explicativa (Sharma, 2013).

Un ejemplo lo tenemos en la actual crisis causada por la COVID-19, ya que organismos nacionales e internacionales, así como los medios de comunicación, publican a diario informes estadísticos, acompañados con gráficos dinámicos e interactivos sobre diversas variables, tales como incidencia, porcentaje de población de diferente edad infectada o vacunada, tanto a nivel nacional, como local o internacional. Son muchas las decisiones que se toman en base a estos datos y que el ciudadano debe aceptar para colaborar en la desaparición de la pandemia. Por ejemplo, en el periodo más duro del confinamiento fue necesaria la sustitución de la docencia presencial por la virtual (Jandrić et al., 2020), situación que ha continuado posteriormente, al menos en forma parcial en muchas universidades.

Por otro lado, en ocasiones esta información tiene una presentación que no siempre es matemáticamente adecuada (Muñiz-Rodríguez, Rodríguez-Muñiz y Alsina, 2020). Consecuentemente, la ciudadanía debe tener un conocimiento estadístico suficiente que le permita alcanzar estas competencias y tener una actitud crítica ante esta información y ante la toma de decisiones en situaciones de incertidumbre (Engel, 2017). Esta competencia no se puede reducir a saber leer las tablas y gráficos estadísticos que encontramos en los medios, sino a ser capaz de interpretarlos de forma adecuada, En concreto, se requiere adquirir un sentido estadístico, que engloba la cultura y razonamiento estadístico, junto con unas buenas actitudes y valoración de la estadística que incremente el interés por su estudio y por su aplicación cuando sea conveniente (Batanero, 2019).

En este trabajo se analiza, en primer lugar, la idea de sentido estadístico y sus componentes. Seguidamente se proponen dos tipos de actividades que el profesor puede organizar en el aula para desarrollarlo en sus estudiantes. En primer lugar, el trabajo con proyectos estadísticos y seguidamente las investigaciones partiendo de noticias basadas en datos estadísticos. Finalmente se resumen las principales conclusiones de la exposición.

## 2. SENTIDO ESTADÍSTICO Y SUS COMPONENTES

En Batanero (2019) se propone la idea de sentido estadístico como la unión de tres componentes. En primer lugar, de una cultura estadística adecuada, que incluye el conocimiento necesario para interpretar

críticamente la información estadística (Gal, 2002). En segundo lugar, se requiere el razonamiento estadístico, que incluye los procesos cognitivos y capacidad de resolver problemas estadísticos. El tercer componente del sentido estadístico lo constituyen las actitudes adecuadas, como la valoración del uso de la estadística en la vida cotidiana, el interés por aprender más estadística y la autovaloración de la propia capacidad para aprender estadística. En lo que sigue comentamos estos componentes, utilizando como contexto la interpretación de uno de los gráficos publicados con información de la COVID (Figura 1).

## 2.1. CULTURA ESTADÍSTICA

Son varios los autores que han analizado la idea de la cultura estadística. Wallman (1993) la definió como la capacidad de comprender y evaluar en forma crítica la información estadística en nuestra vida cotidiana y de valorar la contribución del razonamiento estadístico a la vida personal y profesional. Gal (2002) la define como la capacidad para interpretar y evaluar críticamente información estadística o argumentos que usan esta información en contextos tales como la prensa y el trabajo profesional. Además, supone competencia para utilizar la información estadística cuando sea necesaria para apoyar o comunicar a otros nuestras opiniones. En Batanero (2013) se propone que la cultura estadística requiere el conocimiento de las siguientes ideas estadísticas fundamentales descritas por Burrill y Biehler (2011).

- *Datos.* La importancia de los datos en estadística ha sido resaltada por autores como Moore (1991), que la definió como la ciencia de los datos. Esta importancia ha crecido notablemente con la era del *Big Data*, que requiere de los profesionales nuevas capacidades que no se tienen en cuenta en el currículo escolar (Hardin et al., 2015).
- *Gráficos.* Debido a su presencia en los medios de comunicación e Internet, el aprendizaje de los gráficos estadísticos es una parte esencial de la cultura estadística. Sin embargo, su comprensión no siempre es inmediata y requiere no sólo de conocimientos de los convenios de elaboración, sino de una serie de procesos interpretativos, de cada elemento del gráfico y de todo el gráfico en su conjunto (Friel et al., 2001).
- *Variabilidad.* La variabilidad viene asociada al uso de variables y funciones en todas las ramas de las matemáticas y otras ciencias y aparece en todos los aspectos de nuestra vida y debemos tomar decisiones que dependen de nuestra interpretación de esta variabilidad (Nolan y Provost, 1990), Pero lo que es característico en estadística es la variabilidad aleatoria. Dicha variabilidad aparece en los resultados de los experimentos aleatorios y en el muestreo, en la distribución del estadístico muestral y en los datos recogidos y su comprensión es fundamental para adquirir el concepto más complejo de distribución (Garfield y Ben-Zvi, 2008). La estadística permite buscar explicaciones y causas de la variabilidad, para, a partir de ello, poder hacer predicciones.
- *Distribución.* Aunque el centro de la estadística es la variabilidad, la distribución es el modelo matemático que permite analizarla. Una característica esencial del análisis estadístico es que trata de describir y predecir propiedades de las distribuciones, bien sean distribuciones de datos o distribuciones de probabilidad (Pfannkuch y Reading, 2006). El razonamiento distribucional implica también conectar los datos (distribución de datos), la población de donde se tomaron (distribución de probabilidad) y las posibles muestras de la misma (distribución muestral) (Harradine, Batanero y Rossman, 2011).
- *Asociación y correlación.* En un estudio estadístico generalmente se analiza más de una variable para tratar de determinar sus interrelaciones, lo que lleva a los conceptos de asociación y correlación, como una generalización de la dependencia funcional. La estimación y percepción de la asociación y correlación es una habilidad importante para la toma de decisión, aunque la investigación nos indica que son muchos los sujetos que no la perciben correctamente (Gea et al., 2016).
- *Probabilidad.* La característica principal de la Estadística es hacer uso de modelos aleatorios, a diferencia de otras ramas de la matemática donde se usan modelos deterministas. Ello nos lleva a que el conocimiento de la probabilidad es un requisito necesario para alcanzar una cultura estadística suficiente (Batanero, 2013). Puesto que hay diversas concepciones de la probabilidad, utilizadas dependiendo de las aplicaciones, será necesario adquirir una comprensión suficiente, tanto del enfoque clásico, como del frecuencial y el subjetivo.
- *Muestreo e inferencia.* Relacionar las características de las muestras con las de la población que representan es el principal fin de la estadística. La comprensión del muestreo implica entender las razones que nos llevan a tomar muestras y diferenciar las ideas de representatividad y

variabilidad muestral. Es preciso también adquirir algunas ideas sobre estimación, métodos de muestreo y posibles sesgos en el muestreo.

## 2.2. RAZONAMIENTO ESTADÍSTICO

El sentido estadístico requiere adquirir también los componentes fundamentales del razonamiento estadístico, que, de acuerdo a Wild y Pfannkuch (1999) son los siguientes:

- *Reconocer la necesidad de los datos:* Como se ha indicado, la estadística es la ciencia de los datos, y por tanto los datos son una parte esencial del trabajo estadístico. Es por ello necesario ser capaz de reconocer cuándo se necesitan datos y de qué tipo; por ejemplo, si se necesita realizar un muestreo o se puede estudiar una población completa. También decidir las variables que se precisa analizar y la forma de medirlas o deducirlas (por ejemplo, a través de un cuestionario cuando se realiza un estudio sobre la comprensión de los estudiantes).
- *Percibir la variabilidad:* Ya se ha destacado la importancia de la variabilidad en estadística. Será entonces necesario desarrollar en los estudiantes un sentido de la variabilidad que les permita diferenciar la variabilidad determinista y aleatoria, e identificar las fuentes de variabilidad. Igualmente es necesario el uso de modelos estadísticos que permitan controlar y predecir esta variabilidad.
- *Razonar con modelos estadísticos:* Los estudiantes necesitan diferenciar los datos que forman parte de la muestra (realidad) y el modelo (población), posiblemente descrito por una distribución como la binomial o normal. Se debe fomentar tanto la competencia para generar muestras a partir de un modelo de población, como la de caracterizar el modelo de población y estimar sus parámetros a partir de una muestra. Igualmente se debe fomentar el trabajo con modelos estadísticos como la recta de regresión o las distribuciones básicas, tales como la binomial y normal.
- *Integrar la estadística y el contexto:* Los problemas estadísticos rara vez surgen de la propia estadística o de la matemática, sino de una situación de la vida real, por ejemplo, en economía, gestión, producción, investigación. La modelización supone una observación de la realidad, para simplificarla y trabajar con sus hipótesis y características básicas y construir un modelo matemático a partir de esta simplificación. Una vez que se trabaja con un modelo, la solución se extiende a la situación real, analizando su es o no adecuada. Es importante que el estudiante trabaje todos los pasos del modelo, aunque en la clase en ocasiones se olvida el primer y último paso, es decir, la integración de la estadística con el contexto (Chaput et al., 2011).

## 2.3. ACTITUDES

Hoy día se reconoce el impacto que el dominio afectivo tiene en el aprendizaje de las matemáticas, en especial en la creatividad, la visualización, la intuición o la argumentación (Attard et al., 2016). Este dominio incluye, según Phillip (2007), las emociones, las actitudes y las creencias. Según este autor, las emociones son sentimientos o estados de conciencia, que se distinguen de la cognición, implican sentimientos positivos (por ejemplo, satisfacción) y negativos (por ejemplo, pánico), son transitorios y sirven como fuente para el desarrollo de actitudes.

Di Martino y Zan (2015) sugieren que las actitudes constituyen un puente entre las creencias y las emociones y sus relaciones mutuas. Pueden influir en el comportamiento de la persona con respecto al tema y en su disposición a seguir estudiando o utilizando lo aprendido en el aula y son difíciles de cambiar. Por lo tanto, es importante que los profesores sean capaces de identificar los aspectos afectivos relacionados con su enseñanza o con el trabajo de los alumnos en el aula. Las actitudes hacia un tema (en este caso la correlación y la regresión) incluyen diferentes componentes, como:

- *Afectiva*: sentimientos sobre el objeto en cuestión, si la persona se siente atraída o no por el estudio del tema.
- *Cognitiva*: la autopercepción de la persona con respecto al objeto, si la persona cree que es capaz de manejar el tema.
- Conductual: la inclinación de la persona a actuar hacia el objeto de la actitud de una manera particular, por ejemplo, utilizar el tema para resolver un problema o estudiar más profundamente el tema.
- *Valor*: apreciación de la utilidad, relevancia y valor del tema en la vida personal o profesional.

Una vez descritos estos componentes del sentido estadístico, pasamos a mostrar ejemplos de actividades que contribuyen a desarrollarlo, En primer lugar, los proyectos estadísticos y seguidamente las investigaciones basadas en temas de actualidad.

## 3. PROYECTOS PARA LA CLASE DE ESTADÍSTICA

Un recurso útil para desarrollar el sentido estadístico de los estudiantes es el trabajo con proyectos estadísticos, recomendado por muchos educadores estadísticos para formar mejor a los ciudadanos con conocimientos estadísticos, que hemos analizado en trabajos anteriores (Batanero y Díaz, 2011). Porciúncula y Samá (2014) indican que es un método didáctico que promueve la construcción del conocimiento frente a la memorización y resaltan el hecho de que estos proyectos permiten relacionar diversas áreas escolares y promueven la reflexión de los estudiantes sobre diferentes conceptos estadísticos y su aplicabilidad.

MacGillivray y Pereira Mendoza (2011) indican que, en lugar de introducir los conceptos y técnicas descontextualizadas, en un proyecto se presenta al estudiante las diferentes fases de una investigación estadística: planteamiento de un problema, decisión sobre los datos a recoger, recogida y análisis de datos y obtención de conclusiones sobre el problema planteado. Contribuyen a que el estudiante explore los datos y encuentre en ellos características no esperadas inicialmente y ejerciten tanto la modelización, como el ciclo de investigación propuesto por Wild y Pfannkuch (1999) como parte del razonamiento estadístico.

A continuación, se comenta un posible proyecto adecuado para estudiantes de Educación Secundaria Obligatoria. Se trata de determinar cuál sería el alumno más típico de la clase, considerando diferentes variables. Se preparará una lista de las variables que queremos incluir en el estudio, analizando las diferentes formas en que podrían obtenerse los datos:

- Por simple observación: como el sexo, color de pelo y ojos, si el alumno usa o no gafas.
- Se requiere una medición: como el peso, talla, perímetro de cintura, anchura de hombros o longitud de brazos extendidos.
- Es necesario preguntar a los alumnos: es decir realizar una pequeña encuesta: cuánto deporte practica, número del calzado, cuantas horas duerme, etc.

Una posible tabla de datos de 60 estudiantes se reproduce en la Figura 1, junto con algunos gráficos elaborados con el software CODAP, disponible libremente en Internet en [CODAP - Common Online Data Analysis Platform (concord.org)](). Estos datos pueden grabarse en Excel en formato CSV y el programa lo acepta. En la Figura 1 se muestra la tabla de datos, en la que se han incluido variables nominales, discretas y continuas, incluyendo algunas distribuciones simétricas y asimétricas.

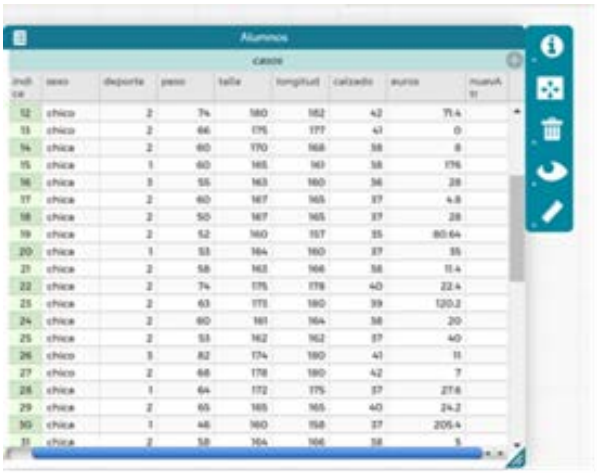



Figura 1: Ejemplo de tabla de datos

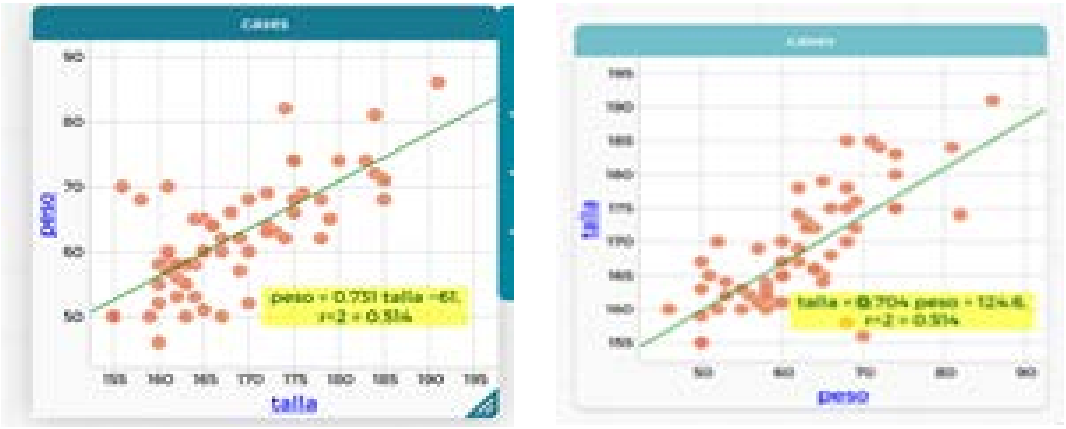
A partir de ellos se puede comenzar el estudio univariante, representando gráficamente las diferentes variables y calculando las medidas de posición central y dispersión y analizando su simetría. Podrían

discutir si hay algún valor atípico para alguna de estas variables y en ese caso, por qué sería preferible la mediana a la media, como medida de posición central. Cada estudiante podría situarse respecto a cada una de estas medidas, calculando en que percentil se situaría.

Las distribuciones obtenidas de medidas físicas serán muy dispares en chicos y chicas por lo que conviene estudiarlas por separado. y algunos gráficos, por ejemplo, la talla y el dinero (Figura 2) que cada uno llevaba en clase ese día y ver en qué variables se aprecian diferencias. Si los estudiantes tienen conocimientos de inferencia, podrían analizarse si las diferencias son estadísticamente significativas.


Figura 2: Representación gráfica de algunas variables en función del sexo.

Es interesante igualmente relacionar las variables cuantitativas mediante el estudio de la regresión. CODAP permite muy fácilmente representar los diagramas de dispersión y calcular la recta de mínimos cuadrados, proporcionando además el valor del cuadrado del coeficiente de correlación (Figura 3). Cambiando una variable por otra, los estudiantes pueden apreciar que hay dos rectas de regresión diferentes, pues uno de los errores más frecuentes en el estudio de la regresión es despejar una de las variables para obtener la recta de regresión de Y sobre X.




Peso en función de la talla Talla en función del peso

Figura 3: Dos rectas de regresión.

Se puede explorar el significado del coeficiente de correlación al cuadrado como medida de la bondad de ajuste analizando diferentes rectas de regresión, por ejemplo, comparando la regresión del peso sobre la talla (Figura 3) o con la regresión de la longitud de brazos sobre la talla (Figura 4). Esta última tiene mejor ajuste explicando la regresión el 75% de la varianza de la longitud de brazos, mientras que del peso solo explica el 51%.

Otra investigación que podría proponer el profesor es sobre el significado de la recta de regresión o de mejor ajuste. CODAP proporciona la posibilidad de utilizar una recta móvil que se sitúa sobre el diagrama de dispersión y el alumno puede mover con el ratón. A la vez se puede calcular la suma de los cuadrados de los residuos (Figura 4). Se trataría de obtener la recta que haga mínima esta suma y una vez el estudiante se aproxima todo lo que puede comparar con la recta de regresión obtenida anteriormente.
Otra discusión posible es por qué se minimiza la suma de cuadrados de los residuos y no la suma de residuos y la relación entre la idea de residuo y la idea de distancia de cada punto a la recta.

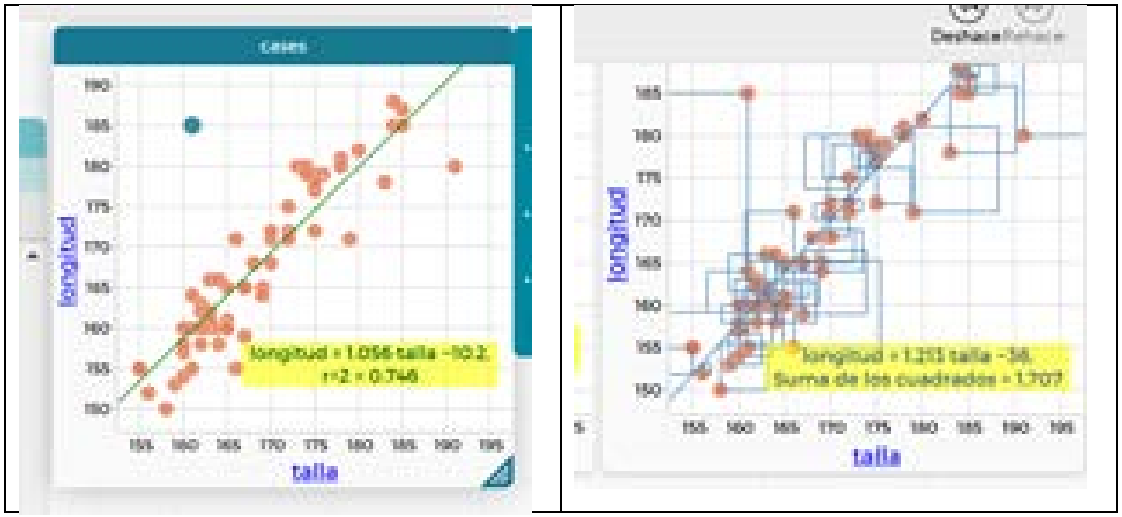



Figura 4: Rectas de ajuste dinámicas.

Estas son algunas de las posibles preguntas que el profesor puede plantear sobre este proyecto, pero lo ideal es que los mismos estudiantes formulen y resuelvan sus propias preguntas e incluso propongan sus propios proyectos. En la plataforma de CODAP hay muchos ficheros de datos disponibles para el trabajo con proyectos.

Existen muchos ejemplos de este tipo de proyectos, ya que diferentes instituciones organizan concursos de proyectos estadísticos para los centros escolares. Algunos ejemplos son el premio al mejor proyecto cooperativo (http://iase-web.org/islp/) promovido por la Asociación Internacional de Educación Estadística como parte del Proyecto Internacional de Alfabetización Estadística o el concurso incubadora de sondeos organizado por la Sociedad Española de Estadística e Investigación Operativa (http://iase-web.org).

El trabajo con proyectos desarrolla la cultura estadística, que, como sugiere Gal (2019), requiere nuevas competencias pues, la comprensión, interpretación y reacción frente a la información estadística no sólo requiere conocimiento estadístico o matemático, sino también habilidades lingüísticas, conocimiento del contexto, capacidad para plantear preguntas, y una postura crítica que se apoya en un conjunto de creencias y actitudes. Por otra parte, el trabajo de los alumnos con los proyectos permite aplicar algunas orientaciones curriculares para la enseñanza de la estadística que sugieren que el alumno sea capaz de abordar problemas de la vida real, organizando y codificando datos, formulando conjeturas sobre los mismos, seleccionando estrategias, y utilizando herramientas y modos de argumentación propios de las matemáticas para resolver problemas y realizar investigaciones.

## 4. INVESTIGACIONES PARA LA CLASE DE ESTADÍSTICA

En general, la enseñanza tradicional de la estadística se ha basado en la resolución de tareas cerradas y bien estructuradas en contextos hipotéticos y controlados, con datos debidamente seleccionados y preparados, y en su mayoría no reales, para la aplicación directa de fórmulas para el cálculo de parámetros estadísticos. En la realidad, los estudios estadísticos suelen ser problemas no estructurados y a menudo ambiguos (Makar y Fielding-Wells, 2011). Si nuestro objetivo principal es conseguir que los alumnos alcancen un pleno sentido estadístico es necesario cambiar la manera de actuar en el aula. Junto con la realización de proyectos, comentada en la sección anterior, podemos integrar en el proceso de enseñanza algunas investigaciones básicas desarrolladas en situaciones reales y cercanas al alumno, a sus

gustos e inquietudes, de tal modo que se trabajen las tres componentes del sentido estadístico: la cultura estadística, el razonamiento estadístico y las actitudes hacia la misma.

Las investigaciones estadísticas reproducen la labor de los estadísticos profesionales, centrada en la resolución de problemas reales que ayudan a fomentar el pensamiento crítico (Wild y Pfannkuch, 1999). Son un campo de conocimiento integrado de conocimientos, procedimientos, habilidades y actitudes para entender y participar críticamente en el mundo (Zapata-Cardona, 2016). En la literatura existen algunos casos de puesta en práctica de investigaciones estadística en las aulas. Por ejemplo, en (Makar y Fielding-Wells, 2011) se utiliza esta metodología para hacer un estudio sobre qué elementos de un avión de papel intervienen en mayor medida a la hora de la distribución de fuerzas durante el vuelo y cuál sería el mejor diseño de entre unos cuantos modelos. En (Zapata-Cardona, 2016) se ofrecen tres casos diferentes de investigaciones: sobre dependencia de los alumnos con dispositivos móviles, sobre igualdad de genero y sobre obesidad infantil.

En este trabajo se plantea una posible propuesta para llevar a cabo diversas investigaciones en torno al uso de gráficos estadísticos, ya que son objetos que necesitan de procesos complejos para su correcta interpretación e involucran a otras partes de la estadística y de la matemática (Arteaga et al., 2018). La búsqueda previa para la elección de gráficos se realiza entre los que aparecen en medios de comunicación y redes sociales tanto en noticias como en espacios publicitarios o mensajes virales, ya que son los que con más seguridad van a proporcionar una temática y un contexto más cercano al discente, añadiendo un factor motivador a la propuesta (Monteiro y Ainley, 2010).

En nuestro caso, recomendamos elegir gráficos que contengan errores o inexactitudes en su construcción que puedan inducir a una interpretación errónea o sesgada de la información si no se posee un adecuado sentido estadístico. La utilización de este tipo de diagramas también tiene un efecto motivador en el alumno y les plantea la necesidad de realizar una lectura crítica de las noticias e informaciones que les llegan (Garzón-Guerrero y Jiménez Castro, 2021).

Para la descripción general de nuestra propuesta nos basaremos en las etapas de indagación estadística propuestas por Wild y Pfannkuch (1999): generación y planteamiento del problema, planificación y recolección de datos, análisis y conclusiones.

- *Generación del problema*: se parte de un fenómeno o situación cotidiana que puede venir en forma de noticia o elemento publicitario conteniendo uno o varios gráficos estadísticos y que proviene de un periódico, revista, TV o alguna de las redes sociales que los alumnos suelen consultar. Los temas seleccionados serán de interés para los alumnos y para su formación personal y social. Es importante no aislar el gráfico del resto, ya que el contexto es importante a la hora de interpretarlo (Monteiro y Ainley, 2010).
- *Estudio del fenómeno*: a continuación, se plantea una puesta en común en la que se cuestiona acerca de la información relevante que son capaces de extraer en una primera lectura del material aportado. Posteriormente se les pedirá analizar el gráfico en mayor profundidad guiándoles por las componentes de razonamiento gráfico de Friel et al. (2001): reconocer componentes estructurales del diagrama, uso de esos elementos en la información representada (incluyendo el mal uso de ellos), contextualizar y traducir entre datos y gráfico, y la capacidad para elegir el gráfico más adecuado. Se insistirá sobre todo en los aspectos de detección de errores o elementos mal construidos en el gráfico y en la posibilidad de usar otro tipo de gráfico más adecuado que el proporcionado.
- *Planteamiento del problema y selección de datos*: a partir del conocimiento de los contenidos de la noticia y su contexto, se originará en los alumnos la necesidad de resolver una situación relacionada de tipo abierto en la que sea necesaria la exploración y tratamiento de datos estadísticos, ya sea por recogida *in situ* o por búsqueda en Internet en alguna de las bases de datos disponibles. El docente será el encargado de orientar a los alumnos, mediante las cuestiones convenientes, para la adecuada definición y planteamiento del problema.
- *Análisis y conclusiones*: se manipularían los datos mediante un software estadístico para obtener los resúmenes estadísticos necesarios y construir gráficos que representen la información encontrada. Daríamos por concluida la actividad de investigación estadística cuando el estudiante comprenda el contexto global, la información mostrada en la noticia y las posibles implicaciones que puede haber en su vida cotidiana, usando su razonamiento crítico para establecer conclusiones fundamentadas, ya sea para entender el fenómeno o para realizar un cambio en su estado actual de actuación y comportamiento.

De esta manera, nuestra propuesta contiene los elementos necesarios para el correcto desarrollo del sentido estadístico: es obligatoria una adecuada cultura estadística que permita conocer los elementos de los gráficos, los datos representados y su distribución…; el razonamiento estadístico es imprescindible para poder razonar con modelos estadísticos e integrar estadística y contexto; y por último, la elección de temas cercanos al alumno y contextos reales influye positivamente en la mejora de la actitud del alumnado.

A continuación, procedemos a mostrar algunos ejemplos de este tipo de investigaciones que pueden ser llevadas a cabo en el aula. Están realizadas para un nivel de Enseñanza Secundaria Obligatoria o Bachillerato, eligiendo los cuatro tipos de gráficos más usuales en los medios de comunicación (líneas, barras, sectores y pictogramas), que también se estudian en este nivel educativo, como punto de partida para las investigaciones.

## 4.1. GRÁFICOS DE LÍNEAS

Son muy empleados en los medios y redes sociales puesto que reflejan claramente la evolución de una variable en el tiempo. Aunque son fáciles de interpretar, su comprensión se dificulta cuando aparece más de una variable, como en el ejemplo mostrado en la Figura 5. A partir del ejemplo se proponen los siguientes pasos:

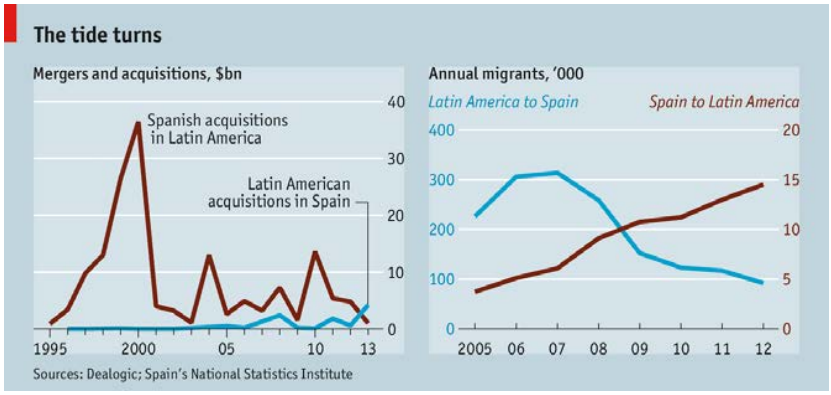



Figura 5: Gráficos de líneas en una noticia sobre migraciones (Fuente: *The Economist*, 25 de enero de 2014).

- *Generación del problema*: Lectura de la noticia, seleccionada del diario estadounidense *The Economist* titulada “Shoe on the other foot”.
- *Estudio del fenómeno*: Discusión con los estudiantes de la noticia, que atiende al cambio en las inversiones económicas y migraciones entre Latinoamérica a España, mostrándose que en los últimos años las inversiones españolas en América Latina han bajado mientras que aumentan en el sentido contrario. La inmigración tiene una tendencia inversa: la migración hacia España ha disminuido, creciendo la contraria. La información se ilustra con dos gráficos de líneas multivariables (Figura 5). Uno de los problemas del artículo es que el gráfico de la derecha puede plantear algunas dudas, pues, a simple vista puede parecer que a partir de 2008 el flujo migratorio se invirtió y van más personas desde España a América que al contrario. Si embargo esto no es cierto, porque en dicho gráfico se representan tres variables: los migrantes en un sentido y en otro y el eje temporal. Las variables de migración tienen escalas diferentes y mientras la primera marca horizontal corresponde a 5000 migrantes al año desde España, por el otro lado marca 100000 migrantes que hacen el viaje opuesto.
- *Planteamiento del problema y selección de datos*: El análisis puede complementarse con varios temas interesantes relacionados con las migraciones. Por ejemplo, podemos preguntar por la tendencia en los últimos años de personas que llegan a España y su país de procedencia o cuáles son los destinos preferidos por los españoles para emigrar al extranjero. Los datos pueden encontrarse abiertos y accesibles desde el Instituto Nacional de Estadística (www.ine.es).
- *Análisis y conclusiones*: los datos serán tratados mediante software estadístico libre como CODAP o tipo hojas de cálculo como LibreOffice.

## 4.2. GRÁFICOS DE BARRAS

Junto con los de líneas, se encuentran entre los más usados en distintos medios, por ser de fácil construcción y lectura. Son útiles para relacionar dos variables donde una de ellas es cualitativa. Se propone un estudio sobre caída del PIB como consecuencia de la pandemia (Figura 6).

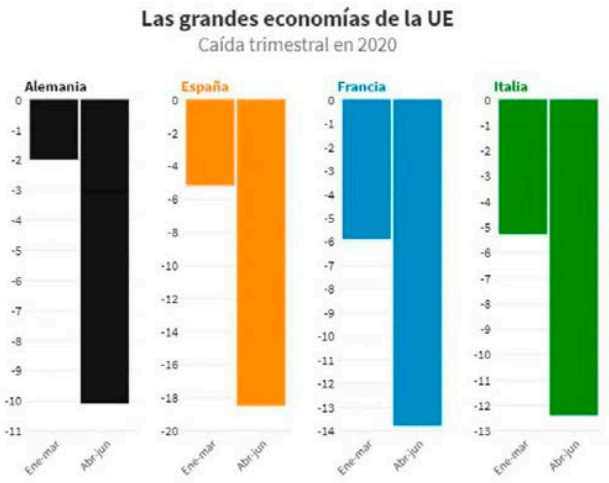



Figura 6: Gráficos de barras relacionados con la caída del PIB por el coronavirus (Fuente: Twitter @rtve, 31 de julio 2020).

- *Generación del problema*: Se proporcionaría a los estudiantes un tweet de la cuenta de RTVE titulada “Caídas históricas del PIB por el coronavirus”.
- *Estudio del fenómeno*: Se leería el texto del tweet, que recoge que el Producto Interior Bruto (PIB) de Alemania, España, Francia y Portugal descendió bruscamente tras el comienzo de la crisis del COVID19. Dicho texto viene acompañado de un gráfico de barras simple (Figura 6) en el que se compara el PIB del primer trimestre del año con el segundo, para los cuatro países mencionados. Se haría notar es que la escala numérica de los diagramas no tiene unidades, lo que puede dificultar la interpretación a quién no esté acostumbrado a temas económicos donde casi todo se mide en valores porcentuales. En un primer vistazo al gráfico puede parecer que la caída es de la misma magnitud en todos los países, pero el diagrama de cada país tiene una escala diferente. En realidad, las mayores caídas se han producido en España (más de un 18%), Francia (casi un 14%), Italia (12%) y Alemania (10%).
- *Planteamiento del problema y selección de datos*: Una posible situación de investigación comenzaría por una definición práctica de lo que es el PIB y cómo se calcula, incidiendo en la diferencia del propio PIB y su tasa de crecimiento y qué nos muestra cada una. Incluiría una búsqueda por la red (por ejemplo, en www.ine.es) de datos de PIB per cápita y de la población parada para su provincia o región para que extraigan posibles correlaciones y conclusiones. Otra posible variante de investigación es la búsqueda de las mayores caídas del PIB en España y qué eventos históricos se asocian a ellas.
- *Análisis y conclusiones*: los datos serán tratados mediante software estadístico libre como CODAP o tipo hojas de cálculo como LibreOffice.

## 4.3. GRÁFICOS DE SECTORES

Los gráficos circulares muestran clases o grupos de datos en proporción al conjunto de datos completo. El círculo completo representa todos los datos, mientras que cada sector o segmento representa una clase o grupo diferente dentro del todo. El número de categorías puede variar, pero generalmente es pequeño. En este caso se utiliza como parte de un anuncio.

- *Generación del problema:* En esta ocasión se ha seleccionado un anuncio publicitario de coches para originar el debate, que se proporciona al estudiante en idioma inglés.

- *Estudio del fenómeno:* La publicidad sigue unas reglas diferentes a la prensa, y no interesa tanto la objetividad del hecho sino vender un producto, en este caso es un coche de una marca coreana de nombre "Verna". El texto que acompaña al anuncio indica que es "el número 1 de ventas en su segmento". La información viene acompañada de un gráfico de sectores tridimensional (Figura 8). Lo primero en lo que nos fijamos es que, efectivamente, el modelo "Verna" es el más vendido, pero el gráfico parece dar a entender que posee casi el 50% del mercado cuando en realidad es el 31%. Por efecto del 3D y de la perspectiva, al estar en primer plano el sector amarillo está distorsionado, destaca más y parece de mayor tamaño de lo que realmente es. Por otro lado, las cifras que aparecen y conforman cada una de las categorías corresponden a las ventas acumuladas, con lo cual no se conoce el tiempo que lleva cada vehículo en el mercado y si la comparación es razonable. Se pediría al alumnado que construyeran un gráfico más adecuado.


Figura 7: Gráfico de sectores usado en publicidad (Fuente: www.chandoo.org, 12 de abril de 2013).

- *Planteamiento del problema y selección de datos:* Se puede partir de la cuestión de cuál es la marca de coche más habitual entre los alumnos de la clase y si esa distribución se mantiene en el resto del país o diferentes regiones. Para estudiarlo, deben realizar una recogida de datos entre todos los compañeros (extensible a otras clases) y comparar posteriormente con datos de venta recogidos de Internet.
- *Análisis y conclusiones:* Los datos serán tratados mediante software estadístico libre como CODAP o tipo hojas de cálculo como LibreOffice.

## 4.4. PICTOGRAMAS

Los pictogramas son utilizados a menudo en la prensa y en la publicidad, debido a que estéticamente se pueden construir con los dibujos o figuras que sean más afines al tema tratado. Puede haber dos tipos: numérico o de barras (formado por figuras de igual tamaño cuyo número es proporcional a la frecuencia representada) o de escala, en el que el tamaño, área o volumen (como en la Figura 8), de cada figura representan la frecuencia de la variable


Figura 8: Pictograma utilizado en un anuncio televisivo de una compañía de telecomunicaciones (Fuente: captura de los autores del vídeo emitido en TV, 2008).

- *Generación del problema*: De nuevo utilizamos un anuncio publicitario como punto de partida, pero esta vez en forma de videoclip emitido por televisión por una compañía de telecomunicaciones.
- *Estudio del fenómeno*: El anuncio de una empresa de telefonía móvil española (representado de color azul en el vídeo) que compara la cantidad de sus clientes con las de dos de sus grandes rivales, en color rojo y naranja. Para ilustrar esas diferencias aparece en el vídeo un gráfico en 3D animado, con pirámides compuestas de teléfonos móviles para representar cada cantidad (Figura 8). Cada pirámide va acompañada del valor numérico que es representado. La compañía azul es claramente la que tiene más clientes, en particular 7 millones más que la siguiente. El problema es que el gráfico no representa esa proporción: la cantidad de clientes está representada por la altura de las pirámides y no por su volumen, como correspondería a un pictograma. De este modo, la compañía azul parece que tiene casi cuatro veces más clientes del valor real.
- *Planteamiento del problema y selección de datos*: Se planteará la pregunta de qué compañías de telefonía tienen los estudiantes y cuál es la más contratada. Otra posible investigación consistiría en sugerir a los alumnos que accedan a las estadísticas de su móvil para comprobar cuanto tiempo utilizan el dispositivo y compartirlo con los compañeros como parte del estudio. El informe final les permitirá saber qué aplicaciones son las más usadas por los estudiantes.
- *Análisis y conclusiones*: los datos serán tratados mediante software estadístico libre como CODAP o tipo hojas de cálculo como las de LibreOffice.

## 5. CONCLUSIONES

El desarrollo de un sentido estadístico adecuado es fundamental para la adquisición del espíritu crítico de los estudiantes. En este trabajo se han definido y analizado las componentes que forman parte del sentido estadístico, unión de la cultura estadística, el razonamiento estadístico y unas actitudes positivas hacia la materia. Para alcanzar el pleno sentido estadístico se han propuesto ejemplos prácticos de proyectos e investigaciones estadísticas que pueden ser implementados en las aulas de Educación Secundaria Obligatoria o Bachillerato. Esperamos que el análisis de los ejemplos presentados pueda motivar a los profesores a buscar otros e introducir los proyectos e investigaciones en la clase de estadística.

## REFERENCIAS

Arteaga, P., Díaz-Levicoy, D., y Batanero, C. (2018). Investigaciones sobre gráficos estadísticos en Educación Primaria: Revisión de la literatura. Revista digital Matemática, Educación e Internet, 18(1), 1-12.

Batanero, C. (2013). Sentido estadístico: Componentes y desarrollo. Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria, 1, 55-61.

Batanero, C. (2019). Statistical sense in the information society. En K. O. Villalba-Condori, A. Adúriz-Bravo, F. J. García-Peñalvo y J. Lavonen (Eds.), Proceedings of the Congreso Internacional Sobre Educación y Tecnología en Ciencias – CISETC (pp. 28-38). Aachen, Germany: CEUR-WS.org.

Batanero, C. y Díaz, C. (2011). Estadística con proyectos. Granada: Universidad de Granada.

Burrill, G. y Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. En C. Batanero, G. Burrill y C. Reading (Eds.), Teaching statistics in school mathematics. Challenges for teaching and teacher education (pp. 57-69). Dordrecht: Springer.

Friel, S.N., Curcio, F.R., y Bright, G.W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. Journal for Research in mathematics Education, 32(2), 124-158.

Chaput, B., Girard, J.C. y Henry, M. (2011). Frequentist approach: Modelling and simulation in statistics and probability teaching. In Teaching Statistics in school mathematics-Challenges for teaching and teacher education (pp. 85-95). Springer, Dordrecht.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science. Statistics Education Research Journal, 16(1), 44-49.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities (with discussion). International Statistical Review, 70(1), 1-51.

Gal, I. (2019). Understanding statistical literacy: About knowledge of contexts and models. Actas del *Tercer Congreso Internacional Virtual de Educación Estadística*. Disponible en www.ugr.es/local/fqm126/civeest.htm.

Garfield, J. y Ben-Zvi, D. (2008). Developing students' statistical reasoning: Connecting research and teaching practice. New York: Springer.

Garzón-Guerrero, J.A., y Jiménez Castro, M. (2021). Un estudio exploratorio de la competencia gráfica de futuros profesores de Portugal e Italia a través de la interpretación de diagramas estadísticos de barras y sectores extraídos de la prensa escrita. Números. Revista de Didáctica de las Matemáticas, 106, 33-42.

Gea, M. M., Batanero, C., López-Martín, M.M. y Arteaga, P. (2016). Research on the perception and learning of correlation and regression. BEIO, Boletín de Estadística e Investigación Operativa, 32(3), 234-256.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P. Peng, R., Roback, P., Temple, D. y Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data". The American Statistician, 69(4), 343-353. https://doi.org/10.1080/00031305.2015.1077729.

Harradine, A., Batanero, C. y Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In Teaching statistics in school mathematics-challenges for teaching and teacher education (pp. 235-246). Springer, Dordrecht.

Jandrić, P., Hayes, D., Truelove, I., Levinson, P., Mayo, P., Ryberg, T. y Hayes, S. (2020). Teaching in the age of Covid-19. Postdigital Science and Education, 2(3), 1069-1230.

MacGillivray, H. y Pereira-Mendoza, L. (2011). Teaching statistical thinking through investigative projects. In Teaching statistics in school mathematics-challenges for teaching and teacher education (pp. 109-120). Springer, Dordrecht.

Makar, K., y Fielding-Wells, J. (2011). Teaching Teachers to Teach Statistical Investigations (pp. 347-358). https://doi.org/10.1007/978-94-007-1131-0_33.

Monteiro, C.E.F., y Ainley, J.M. (2010). The interpretation of graphs: Reflecting on contextual aspects. Alexandria: Revista de Educação em Ciência e Tecnologia, 3(2), 17-30.

Moore, D.S. (1991). Teaching statistics as a respectable subject. En F. Gordon y S. Gordon (eds.), Statistics for the twenty-first century (pp. 14-25). Mathematical Association of America.

Muñiz-Rodríguez, L., Rodríguez-Muñiz, L.J. y Alsina, Á. (2020). Deficits in the statistical and probabilistic literacy of citizens: effects in a world in crisis. Mathematics, 8(11), 1872; https://doi.org/10.3390/math8111872.

Nolan, T.W. y Provost, L.P. (1990). Understanding variation. *Quality Progress, 23*(5), 70-78.

Pfannkuch, M. y Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal., 5* (2), 4-9.

Porciuncula, M. y Samá, S. S. (2014). Teaching Statistics through Learning Projects. Statistics Education Research Journal, 13(2). 177-186.

Ridgway, J. (2016). Implications of the data revolution for statistics education. International Statistical Review, 84(3), 528-549.

Sharma, S. (2013). Assessıng students' understandıng of tables and graphs: implıcatıons for teachıng and research. International Journal of Educational Research and Technology, 4(4), 61-69.

Wild, C.J., y Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. International Statistical Review, 67(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x.

Zapata-Cardona, L. (2016). Enseñanza de la estadística desde una perspectiva crítica. Yupana. Revista de Educación Matemática de la UNL, 10, 30-39.

# Curso TIC

*XV Congreso Galego de Estatística e Investigación de Operacións*
*Santiago de Compostela, 4, 5 e 6 de novembro de 2021*

# CURSO: APRENDENDO FERRAMENTAS TIC PARA O ENSINO DA ESTATÍSTICA

María Isabel Borrajo García, Mercedes Conde Amboage e María José Ginzo Villamayor

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

## RESUMO

Este curso pretende dotar ao profesorado dos distintos niveis educativos dunha ampla oferta de ferramentas útiles para o ensino da Estatística. Para isto, elaborouse un programa que comprende dende a folla de cálculo ata presentacións dinámicas, pasando por R, LaTeX ou sistemas para a aleatorización de exames, entre outros. Ademais, o curso terá unha sesión final presencial na que se introducirá aos asistentes na creación de vídeos, partindo da elaboración dos brutos, a súa edición, a posterior exportación e a publicación do resultado final. Nesta sesión proporcionaranse, ademais, unha serie de ideas e consellos sinxelos de cara a obter o mellor produto final posible con medios que están ao alcance de todos/as.

**Palabras e frases chave**: ferramentas TIC; ensino da Estatística; *software* libre.

## 1. MOTIVACIÓN

Poderiamos dicir que as ferramentas TIC (Tecnoloxías da Información e da Comunicación) son ese conxunto de tecnoloxías que foron desenvoltas co propósito de manexar a información e comunicala dun lugar a outro. As ferramentas TIC son xa parte das nosas vidas, posto que as empregamos para case calquera actividade: *Alexa* pode acender unha lámpada ou os *smartwatches* poden medir o noso ritmo cardíaco. En ámbitos profesionais, poderiamos dicir que se empregan en practicamente todos os sectores, entre eles na educación, que é o que nos interesa neste curso.

Nos últimos anos, e especialmente coa pandemia, que promoveu a educación telemática por mor dos confinamentos, universidades, escolas de negocios, colexios e institutos, víronse na obriga de implantar as TIC con campus virtuais, grupos específicos en Internet, uso de aplicacións e plataformas como *Microsoft Teams* ou *Moodle*, ... As vantaxes son indiscutibles e ben recoñecidas tanto polo alumnado como polo profesorado, se ben é certo que este último sector puido verse un pouco desbordado ante uns requirimentos e necesidades para os que non se sentían suficientemente formados.

Neste curso preténdese transmitir a profesorado dos diferentes niveis educativos as posibilidades existentes en materia de ferramentas TIC para o ensino da Estatística. A Estatística, non é unha rama das Matemáticas, senón que constitúe unha ciencia en si mesma (a ciencia dos datos), cun carácter singular acreditado por uns obxectivos propios, conceptos distintivos e modos de razoamento específicos. Por todo isto, as ferramentas que se aplican, aínda que en certa medida comúns ás doutras ciencias, deben de estar orientadas ao proceso de ensinanza-aprendizaxe particular da Estatística.

## 2. CONTIDOS

A continuación presentamos un esquema dos 8 bloques temáticos nos que se dividen este curso; en xeral, cada un deles correspóndese coa explicación e ilustración dunha única ferramenta, salvo no caso do módulo 5, que contempla dúas ferramentas cun obxectivo común de levar a cabo tarefas de avaliación, e do módulo 7, que inclúe diversas ferramentas máis sinxelas que poden ser empregadas na aula con finalidades moi variadas.

- Módulo 1. Algunhas pinceladas da folla de cálculo.
  Neste bloque aprenderemos a usar a folla de cálculo. Faremos un percorrido comezando nas funcións máis sinxelas, pasando por algunhas funcións comúns, pola elaboración de táboas dinámicas e mesmo pola creación de novas funcións segundo as nosas necesidades.

- Módulo 2. Primeiros pasos co programa estatístico R.
  Neste bloque presentaremos o programa estatístico R. Aprenderemos tanto o máis básico como é a súa instalación ou o uso do mesmo como calculadora, como os tipos de obxectos, librarías útiles, lectura e escritura de datos, así como as principais ferramentas que nos permiten facer unha análise de estatística descritiva dunha mostra.

- Módulo 3. O editor de textos científicos LaTeX.
  Neste bloque instalaremos e aprenderemos a usar o que probablemente sexa o editor de textos científicos por excelencia: LaTeX. Veremos como elaborar un documento (xa sexa de tipo artigo ou de tipo libro), como incluír imaxes, listas, símbolos matemáticos, ecuacións, matrices, táboas, ...

- Módulo 4. Introdución a R Markdown.
  Neste bloque introduciremos a linguaxe R Markdown. Trátase dunha linguaxe moi lixeira que nos permite elaborar informes ou documentos que conteñan código de R inserido. Para isto empregaremos a interface gráfica RStudio e aprenderemos os elementos básicos para crear un documento a partir de cero: cabeceira, anacos de código de R, inserción de elementos como imaxes ou táboas, ...

- Módulo 5. Avaliación: aleatorización de exames e elaboración de cuestionarios.
  Neste bloque sacaremos proveito das ferramentas TIC para o proceso de avaliación. Aprenderemos a elaborar o que denominaremos exames aleatorios, é dicir, exames individualizados para cada alumna/o pero mantendo unha estrutura e complexidade común. Tamén veremos como crear, do xeito máis eficiente posible, cuestionarios en Moodle.

- Módulo 6. Presentacións dinámicas.
  As presentacións dinámicas son un elemento máis que podemos empregar para captar a atención das/os nosas/os interlocutoras/es. Neste bloque abordamos a creación deste tipo de documentos con RStudio, obtendo un documento final en HTML5 que se pode abrir en calquera navegador.

- Módulo 7. Outras ferramentas útiles.
  Neste bloque imos presentar unha listaxe de ferramentas e aplicacións máis sinxelas que as presentadas anteriormente, pero que nos permitirán levar a cabo concursos ou enquisas, atopar datos ilustrativos, gamificar as nosas aulas, xerar gráficos máis desenfadados, ...

- Módulo 8. Produción e edición de vídeo.
  Este derradeiro bloque temático correspóndese coa sesión presencial que se impartirá durante o XV Congreso SGAPEIO. Nesa sesión aprenderemos os conceptos básicos sobre edición de vídeo empregando ferramentas de *software* libre.

## 3. METODOLOXÍA

O curso ten unha duración de 36 horas, das cales 33 horas realízanse de xeito telemático, mentres que as 3 últimas horas impartirase de xeito presencial durante as datas de celebración do XV Congreso SGAPEIO.

O inicio do curso levouse a cabo a finais de setembro cunha sesión de presentación online a través dun foro de cada un/unha dos/as participantes. Estas/es disporán de apuntamentos e presentacións que os guiarán na súa aprendizaxe, e que están dispoñibles na aula virtual do curso. A idea é que as/os participantes poidan ir lendo e comprendendo eses apuntamentos, á vez que van poñendo en práctica nos seus ordenadores as indicacións que alí se detallan. Tamén dispoñen dun foro por bloque temático para preguntar dúbidas, así como dun calendario que guía o seu progreso na aula virtual.

## 4. AVALIACIÓN

Ao final de cada un dos bloques temáticos, salvo do bloque 8 que se realiza de xeito presencial, proponse unha tarefa a que as/os participantes deben realizar e entregar a través da aula virtual. Esa tarefa require poñer en práctica o aprendido sobre a(s) ferramenta(s) correspondente(s). as/os participantes que queiran obter o diploma de superación do curso, deberán de realizar correctamente o 85% desas tarefas solicitadas, e asistir ademais á sesión final a realizar durante o XV Congreso SGAPEIO.

## REFERENCIAS

CTAN: Comprehensive TEX Archive Network. (2021). LaTeX2e: An unofficial reference manual. http://tug.ctan.org/info/latex2e-help-texinfo/latex2e.pdf

Libre Office: The document foundation (2021). Calc Guide. https://documentation.libreoffice.org/assets/Uploads/Documentation/en/CG7.1/CG71-CalcGuide.pdf

Moodle web page. https://moodle.org/?lang=es

R Studio support. (2021) R Presentations. https://support.rstudio.com/hc/en-us/sections/200130218-R-Presentations

Verzani, J. (2005). Using R for introductory Statistics. Chapman and Hall.

Xie, Y., Allaire, J. J. e Grolemund, G. (2021). R Markdown: The Definitive Guide https://bookdown.org/yihui/rmarkdown/

Zuur, A.F., Ieno, E.N. e Meesters, E.H.W.G. (2009). A Beginner's Guide to R. Springer.

## Organizadores



DEPARTAMENTO DE ESTATÍSTICA,
ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

## Patrocinadores










## Colaboradores