XVI Congreso Galego de Estatística e Investigación de Operacións

I Xornadas Innovación Docente na Estatística e IO

2023

UNIVERSIDADE DA CORUÑA
Departamento de Matemáticas

#SGaPEIO

XVI Congreso Galego de Estatística e Investigación de Operacións

I Xornadas de Innovación Docente na Estatística e Investigación de Operacións

Libro de actas

A Coruña, do 26 ao 28 de outubro de 2023

XVI Congreso Galego de Estatística e Investigación de Operacións
I Xornadas de Innovación Docente na Estatística e Investigación de Operacións

Libro de actas

Elaborado por:
Comité organizador

ISBN: 978-84-09-50202-8

Presentación

O Departamento de Matemáticas da Universidade da Coruña, a petición da Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO), organiza o XVI Congreso Galego de Estatística e Investigación de Operacións. Nesta ocasión, o congreso celébrase na Escola Técnica Superior de Enxeñaría de Camiños, Canais a Portos (ETSECCP) de A Coruña do 26 ao 28 de outubro de 2023. É a terceira vez que a SGAPEIO celebra na nosa cidade o seu congreso bienal, despois de que en 1993 tivese lugar o I Congreso e en 2013 o XI Congreso. Como en edicións anteriores, entre os distintos actos incluiranse conferencias, mesas redondas, obradoiros e comunicacións, ademais de actividades de carácter cultural coas que intentaremos facer máis atractiva a participación. O obxectivo deste congreso é contribuír á divulgación dos avances na teoría, nas aplicacións e na didáctica da Estatística e da Investigación de Operacións.

Ademais, o Departamento de Matemáticas da UDC e a SGAPEIO, en colaboración coa Asociación Galega de Profesorado de Educación Matemática (AGAPEMA), organizan as I Xornadas de Innovación Docente na Estatística e Investigación de Operacións. Esta xornada, dirixida a profesorado de ESO, bacharelato e universidade, celebrarase o venres 27 de outubro pola tarde e o sábado 28 de outubro pola mañá.

O presente libro de actas recolle o resumo das catro conferencias plenarias e os 56 traballos presentados: 41 comunicacións orais, das que 9 son traballos que optan ao premio a investigadores novos e 3 son traballos presentados na sesión de biometría que organizan conxuntamente a SGAPEIO e a Sociedade Portuguesa de Estatística (SPE); 11 pósteres e 4 comunicacións orais nas xornadas de innovación docente. Agradecémoslles aos autores destes traballos as súas contribucións

O Comité organizador

Comité científico

Ricardo Cao Abad (Universidade da Coruña) Presidente
Balbina Casas Méndez (Universidade de Santiago de Compostela)
Ignacio García Jurado (Universidade da Coruña)
Mª Carmen Iglesias Pérez (Universidade de Vigo)
Esther López Vizcaíno (Instituto Galego de Estatística)
Leticia Lorenzo Picado (Universidade de Vigo)
Salvador Naya Fernández (Universidade da Coruña)
Matilde Ríos Fachal (CPI Cruz do Sar)
Jacobo de Uña Álvarez (Universidade de Vigo)
Juan M Vilar Fernández (Universidade da Coruña)

Comité organizador

M Amalia Jácome Pumar (Universidade da Coruña) Presidenta
Ana López Cheda (Universidade da Coruña) Secretaria
Germán Aneiros Pérez (Universidade da Coruña)
María Luisa Carpente Rodríguez (Universidade da Coruña)
Julián Costa Bouzas (Universidade da Coruña)
Graciela Estévez Pérez (Universidade da Coruña)
Mario Francisco Fernández (Universidade da Coruña)
María José Ginzo Villamayor (Universidade de Santiago de Compostela)
Ángel Manuel González Rueda (Universidade de Santiago de Compostela)
Maria Teresa Malheiro (Universidade do Minho)
María Cristina Muiños Roca (Universidade da Coruña)
Manuel Oviedo de la Fuente (Universidade da Coruña)
Beatriz Piñeiro Lamas (Universidade da Coruña)
Manuel Antonio Presedo Quindimil (Universidade da Coruña)
Luis F Rodríguez Fraguela (Universidade da Coruña)

Contidos

Conferencias plenarias

Mesas redondas

Obradoiros

Venres, 27 de outubro

16:00 – 18:00

18:15 – 20:00

Comunicacións orais

Xoves, 26 de outubro

11:30 – 12:45

Premio investigadores novos (modalidade A)

Estatística pública

Métodos de estimación

Venres, 27 de outubro

9:00 – 10:00

Optimización

Aplicacións da estatística I

Contrastes de hipótesis

12:30 – 13:30

Premio investigadores novos (modalidades A e B)

Teoría de xogos

Sesión de biometría: SGAPEIO – SPE

Sábado, 28 de outubro

9:00 – 10:15

Aplicacións da estatística II

Estatística non paramétrica

Pósters

Xoves, 26 de outubro

Xornadas de innovación docente na estatística e investigación de operacións

Sábado, 28 de outubro

9:00 – 10:15

Conferencias plenarias

Navigating disease mapping: Unraveling Models, High Dimensionality and Recent Applications

Lola Ugarte (Universidad Pública de Granada)

In this presentation, we will embark on a journey through the realm of disease mapping, exploring both univariate and multivariate spatio-temporal statistical models.

Some ideas related to the analysis of high-dimensional data in this context will also be presented. Moreover, we will spotlight the latest applications of these models, addressing both violence against women in India and cancer mortality.

Some data science models under the location analysis lens

Justo Puerto (IMUS, Universidad de Sevilla)

The design, management and use of any type of complex network requires a methodology to handle its parameters, detect deficiencies and coordinate their resources to solve the problems that arise. Developing methods to carry out such actions demands, among other things, the preliminary screening of large masses of data, quantitative analysis to design better information structures, often organized as networks, and the solution of optimization problems related to clustering, location, routes, allocation of flows and traffic of any kind, distribution of intelligent sensors, early detection of extreme observations, profiling user behavior and operations planning, often under an environment of risk or uncertainty, etc. All those operations involve large masses of data that must be integrated in all phases of the operational analysis. The standard approach of handling separately/sequentially data and design is defective and lacks the important gain of integration. Data integration, data reduction, feature selection, outliers detection, intelligent segmentation or separability are the driving challenges that relies on tools such as machine learning, statistical analysis, optimization, and mathematical programming.

One step forward to bring the gap of integration in data science is the application of techniques from optimization and statistics. In this talk, we focus on one important challenge: "integration of design, optimization and data". This approach sets a very ambitious objective: to improve the data science paradigm integrating techniques of location and networks analysis and vice versa. The two features of datafication and universalization of the available information establish a subtle difference with the standard methodologies of location science and network analysis: the representation of complex environments with large masses of data imposes the need to apply more advanced tools of mathematics and machine learning to allow the design, the effective treatment and use of data at all levels and the optimization of the problems that arise from them.

In general terms, the challenge that is currently posed in this field is, not only, to incorporate the methodology of data science into the analysis of large scale networks, in order to deal with problems that involve large masses of data ("Big Data"); but also, reciprocally, how to make use of the tools and models of optimization and network design in data science. This talk will surf over recent results of our team (see references) in this respect, showing how modern location analysis techniques improve several machine learning methodologies including regression, unsupervised and supervised classification and community detection.

Modelos de cooperación en problemas de inventario con demanda constante

Gloria Fiestras Janeiro (Universidade de Vigo)

El modelo EOQ es un modelo de gestión de inventario de un único producto con demanda constante por unidad de tiempo al que se enfrenta una empresa. Su objetivo es encontrar el tamaño óptimo de pedido de modo que se minimice una determinada función de coste. Si varias empresas demandan el mismo producto pueden cooperar realizando pedidos conjuntos para minimizar el coste total. En esta situación ha estudiarse si resulta beneficioso esta acción y, en caso afirmativo, establecer cuál será el reparto del coste final soportado por cada empresa. En esta charla revisaremos algunos modelos cooperativos de inventario EOQ con diferentes funciones de coste y expondremos algunos repartos del coste final mostrando sus propiedades.

El espejismo de la mayoría

Eduardo Sáenz de Cabezón (Universidad de La Rioja)

Las matemáticas pueden ayudarnos a entender nuestro comportamiento colectivo, también en redes sociales. La estructura de las redes y los algoritmos que regulan la información que recibimos y enviamos hacen que a veces se den fenómenos inesperados, no siempre positivos. Conocerlos puede ayudarnos a tener una mejor experiencia de acceso a las redes sociales. Y en este empeño, las matemáticas son un aliado insustituible.

Mesas redondas

O despoboamento en Galicia, cara a onde imos?

Moderadora:

Esther López Vizcaíno (IGE)

Participantes:

Isabel del Río Viqueira (IGE)
Inés Santé Riveira (Consellería de Medio Rural)
Javier Sanz Larruga (UDC)
Edelmiro López Iglesias (USC)

Innovación docente e divulgación da estatística e investigación de operacións

Moderadora:

Paula Blanco Mosquera (Xunta de Galicia)

Participantes:

Eduardo Sáenz de Cabezón (Unirioja)
Marta Pérez Rodríguez (UVigo)
Elena Vázquez Abal (USC)
Mª Cristina Naya Riveiro (UDC)
Javier Álvarez Liébana (UCM)

Obradoiros

Te lo cuentan los datos

Marta Pérez Rodríguez, Beatriz Álvarez Díaz

Te lo cuentan las matemáticas é un proxecto de divulgación das matemáticas. Con el preténdese mellorar a educación matemática na última etapa de Educación Primaria e primeiros cursos de Educación Secundaria, dende un punto de vista científico, tecnolóxico, transversal, funcional, histórico e de xénero. Preténdese así contribuír a fomentar a utilidade social e económica das matemáticas, facilitando o achegamento entre as matemáticas e outras ciencias experimentais, humanísticas e sociais, así coma con outros ámbitos culturais, artísticos e sociais. Nesta tarefa é esencial promover actitudes e comportamentos favorables cara ás matemáticas coma unha ferramenta necesaria e omnipresente na vida cotiá e na comprensión do mundo; aumentando así a interese, o gusto e a curiosidade cara elas.

O proxecto desenvólvese en bloques temáticos con talleres e encontros con mozas matemáticas. Un dos bloques, Te lo cuentan los datos, está adicado á probabilidade e á estatística e inclúe dous talleres de título Te lo cuenta una epidemióloga.

A formación nos talleres sobre datos Te lo cuenta una epidemióloga está dirixida a profesorado de Educación Primaria e Secundaria, co obxectivo de que posteriormente poidan implementar os talleres nas súas aulas. A formación consta dun taller manipulativo e dun taller tecnolóxico.

No taller manipulativo preténdese que os participantes adquiran conceptos de probabilidade a través de materiais coma dados, vasos ou moedas. A través da realización de diversos experimentos e da toma de datos asociada, abórdanse os conceptos de probabilidade experimental e probabilidade teórica. A partir deste estudo realízanse predicións aplicadas a xogos de mesa.

No taller tecnolóxico realízase un tratamento dos datos dun partido de baloncesto, co obxectivo de axudar ao adestrador a tomar boas decisións que o axuden a gañar o partido. Para iso, úsase unha folla de cálculo que permite realizar unha análise destes datos utilizando ferramentas de estatística descritiva: medidas estatísticas, representacións gráficas.

Conectando R con outras linguaxes

Laura Davila Pena, José Ameijeiras Alonso

Ás veces, ao programar en R, atopamos unha parte do código que, aínda que é sinxela, non se executa tan rápido coma nos gustaría. Neses casos, é posible que outras linguaxes de programación de medio nivel, coma Fortran ou C, poidan realizar a mesma tarefa nun tempo computacional moito menor. Non obstante, cambiar todo o código a unha linguaxe diferente pode ser un proceso longo e custoso. Nese momento, podemos adoptar unha solución intermedia: combinar o código que xa

temos en R con funcións implementadas de forma eficiente noutras linguaxes. O obxectivo deste taller é ofrecer unha breve introdución sobre como realizar este proceso, proporcionando algunhas ferramentas e nocións básicas para conectar R con outras linguaxes de programación, coma Fortran, C ou C++.

Python e Shiny

Marta Rodríguez Barreiro

Python é unha das linguaxes de programación máis usadas na actualidade, en gran parte debido á súa versatilidade. É unha linguaxe de alto nivel que se pode empregar para diversos propósitos como o densenvolvemento de software, scripts web, ou enxeñería de datos. Está creado baixo unha licencia de código aberto e conta cun amplo catálogo de funcións integradas na súa biblioteca estándar.

Pola súa parte, Shiny é unha librería que permite crear aplicacións interactivas de forma sinxela. Ata agora, Shiny estaba dispoñible só con R, pero na actualidade xa se creou a librería para Python. Desta forma é posible crear aplicacións con visualizacións interactivas empregando unicamente código Python e sen ter ningún coñecemento de programación web.

Neste obradoiro abordaremos unha introdución á programación con Python e empregaremos esta linguaxe para crear aplicacións sinxelas con Shiny que nos permitan ver todo o potencial destas ferramentas.

Ferramentas para traballar a estatística e investigación de operacións na aula

María José Ginzo Villamayor, Alejandro Saavedra Nieves

Neste obradoiro empregaremos ferramentas de libre acceso nas aulas, como Maxima ou R. Revisaremos con estes software os contidos aplicados das materias do ámbito das matemáticas en calquera dos niveis educativos do ensino medio, así como, a súa ilustración con exemplos da vida real.

Comunicacións orais

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Predicting area-level exposure indexes for sex occupational segregation: A Compositional Multivariate Fay-Herriot Model Approach.

Esteban Cabello[1], Domingo Morales[1], Agustín Pérez[2]

[1]Centro de Investigación Operativa, Universidad Miguel Hernández de Elche.
[2] Departamento de Estudos Económicos e Financeiros, Universidad Miguel Hernández de Elche.

ABSTRACT

This paper presents a novel methodology for predicting area-level proportions of employed men and women across various occupation sectors, along with estimating exposure indexes between sex groups. The challenge arises from the compositional nature of the direct estimators of proportions, which tend to be imprecise when sample sizes are small. To overcome this problem, we propose to use a compositional multivariate Fay-Herriot model. By applying log-ratio transformations to the direct estimators of proportions, we can effectively capture the underlying structure and dependencies within the data. Small area estimators for proportions and exposure indexes are derived from the fitted model, and their corresponding root-mean-squared errors are estimated using parametric bootstrap techniques. To demonstrate the applicability of our approach, we conduct a case study using data from the quarters 3 and 4 of the Spanish Labour Force Survey of 2022. The primary objective is to investigate the state of sex occupational segregation in Spanish provinces, thereby providing valuable insights into this socio-economic phenomenon.

Key words: Small area estimation, multivariate Fay-Herriot model, compositional data, bootstrap, exposure index, occupation sectors, Labour Force Survey.

1. INTRODUCTION

The Exposure Index (EI) is a widely used interaction index among the so-called segregation measures. In particular, the EI measures the potential contact between two groups, minority and majority, within common geographical areas (Bell, 1954) and here lies its importance, as it is applicable to study differences between sex, race, culture or religion among others. For this reason, it can be used to measure how distribution by occupation sectors of a interest group of people differs from a second that serves as reference. Our investigation delves into utilizing the EI as a measure to assess the extent of occupational segregation based on sex. In this context, we designate sex as the grouping variable, while the distribution or classification variable is the labour occupation sector. As Massey and Denton (1988) describe, exposure indexes do not attempt to measure segregation as a deviation from an ideal equality (whose role is associated with measures of Evenness), but rather measure the experience of segregation felt by the average member of the minority or majority. The EI measures the extent to which employed women are exposed to employed men in different occupational sectors. Equivalently, EI can be interpreted as the hypothetical probability of a member of a minority group encountering a member of the majority group in the same occupational sector. In order to accomplish this, we compare the relative representation of men and women employed across different occupational sectors and derive a numerical value that is equal to or greater than zero. When dealing with only two groups, the combined sum of the EI equals one, where lower interaction values indicate a higher degree of segregation. However, in scenarios involving more than two groups, the EI do not add up to one. Furthermore, the symmetry property is not satisfied generally in this index, only if the two groups constitute the same proportion of the population. This is an unlikely condition in practice. Bell (1954) describes the basic properties of the EI.

We can estimate EIs by direct methods when we obtain the data from surveys with large samples or from administrative records; that is, in those cases in which the available information is highly reliable. However, the direct estimation of EIs is not accurate in territorial areas or in social groups where there are few data. The introduction of statistical methodologies to address these cases motivate the research problem of this paper. Small Area Estimation (SAE) provides methodologies and statistical techniques to estimate EIs in domains with a small number of sample data. The main idea is to use statistical models to borrow strength from other domains, auxiliary variables, and dependency and correlation structures in hierarchical, spatial, or temporal data. Rao and Molina (2015), Pratesi (2016) and Morales et al. (2021) provide an introduction to this branch of Statistics.

We apply the new statistical tools to data from the Spanish Labour Force Survey (SLFS) of quarters 3 and 4 of 2022. In this sense, we use past recent information (third quarter) to get more accurate and reliable information about the present information. The sociological research target is to study the provincial EIs between the distributions of women and men in the different employment sectors. We investigate statistical procedures to map EIs at different levels of aggregation, with the aim of providing information to policy makers that facilitates the implementation of equality policies. We introduce a statistical methodology that has three novel aspects: (1) we apply log-ratio transformations to aggregated compositional data, (2) we fit MFH models to the transformed data, and (3) we construct model-based predictors of proportions, counts, and EIs for small areas and time periods.

The rest of the paper is organized as follows. Section 2 introduces the probabilistic framework, the EI, the data, and the SAE problem. Section 3 contains the new MFH model, the algorirhm to calculate the residual maximum likelihood estimators (REML) of the model parameters, the predictors of proportions of employed men and women by occupation sectors, the predictors of EIs, and the mean squared error (MSE) bootstrap estimators. Section 4 includes some simulation experiments to investigate the performance of the EI predictors and the effect of the number of bootstrap replicates in the accuracy of the MSE estimators. Section 5 deals with the application to real data. Section 6 provides some relevant conclusions.

2. THE PROBLEM OF INTEREST AND THE DATA

This section introduces the probabilistic framework under which the statistical methodology for SAE is developed and describes the available data. The variable that allows introducing the probability vectors (or distributions) to be compared, and which are the arguments of the EIs, is the occupation sector (OC). This variable has $R = 7$ categories that cover a wide range of job occupations in Spain. The OC categories are

OC1 Directors and managers. Management of companies and public administrations,

OC2 Scientific and intellectual technicians and professionals,

OC3 Military occupations. Technicians and support professionals,

OC4 Accounting, administrative and other office employees,

OC5 Workers in catering services, protection and trade vendors,

OC6 Unskilled workers and skilled workers in agriculture, livestock, forestry and fishing,

OC7 Plant and machinery operators and assemblers. Artisans and skilled workers in the manufacturing, construction and mining industries.

In order to mathematically formulate the EIs between the distributions of labor occupation by sex for each province and period of time, we need to introduce some notation. Consider the variables sex, time, province and occupation sector, with values $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$ and $r = 1, \ldots, R$, respectively. The finite population of interest (employed people) can be divided by sex, $U = U_1 \cup U_2$, $U_1 \cap U_2 = \emptyset$, by time, $U_k = \cup_{t=1}^{2} U_{kt}$, $k = 1, 2$, $U_{kt_1} \cap U_{kt_2} = \emptyset$, $t_1 \neq t_2$, by province, $U_{kt} = \cup_{d=1}^{D} U_{ktd}$, $k = 1, 2$, $t = 1, 2$, $U_{ktd_1} \cap U_{ktd_2} = \emptyset$, $d_1 \neq d_2$, and by sector, $U_{ktd} = \cup_{r=1}^{R} U_{ktdr}$, $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$, $U_{ktdr_1} \cap U_{ktdr_2} = \emptyset$, $r_1 \neq r_2$, where D is the number of provinces and R is the number of occupation sectors. Let N, N_k, N_{kt}, N_{ktd} and N_{ktdr} be the corresponding population sizes. The partition of U by time is $U = \cup_{t=1}^{2} \Omega_t$, $\Omega_1 \cap \Omega_2 = \emptyset$, and the partition of Ω_t by occupation sector is $\Omega_t = \cup_{r=1}^{R} \Omega_{tr}$, $\Omega_{tr_1} \cap \Omega_{tr_2} = \emptyset$, $r_1 \neq r_2$.

For $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$, $r = 1, \ldots, R$, $j = 1, \ldots, N_{ktd}$, the definition of the indicator variables, $z_{ktdj,r}$, of occupation sector r in the sex-time-province U_{ktd}, is $z_{ktdj,r} = 1$ if $u_{ktdj} \in \Omega_r$ and $z_{ktdj,r} = 0$ otherwise, where u_{ktdj} is the j-th individual of U_{ktd}. Let $Z_{ktdr} = \sum_{j \in U_{ktd}} z_{ktdj,r}$ and $\overline{Z}_{ktdr} = Z_{ktdr}/N_{ktd}$ be the total and the proportion of employed people of sex k, time t and province d that works in occupation sector r, respectively. It holds that $\sum_{r=1}^{R} Z_{ktdr} = N_{ktd}$ and $\sum_{r=1}^{R} \overline{Z}_{ktdr} = 1$. We further define the vectors $Z_{ktd} = (Z_{ktd1}, \ldots, Z_{ktdR-1})'$ and $\overline{Z}_{ktd} = (\overline{Z}_{ktd1} \ldots, \overline{Z}_{ktdR-1})'$, $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$.

The EI of province d at time period t, between the OC distributions of men and women, is $E_{td} = E(\overline{Z}_{1td}, \overline{Z}_{2td}) = E(Z_{1td}, Z_{2td})$, where

$$E_{td} = \sum_{r=1}^{R} \overline{Z}_{2tdr} \frac{\overline{Z}_{1tdr}}{\frac{N_{2td}}{N_{1td}} \overline{Z}_{2tdr} + \overline{Z}_{1tdr}} = \frac{1}{N_{2td}} \sum_{r=1}^{R} \frac{Z_{2tdr} Z_{1tdr}}{Z_{2tdr} + Z_{1tdr}}, \quad t = 1, 2,\ d = 1, \ldots, D.$$

This paper uses SLFS data to estimate E_{td}, $t = 1, 2$, $d = 1, \ldots, D$. The SLFS is published quarterly and provides information on the Spanish labor market. It is a survey with a two-stage sampling design. In the first stage, the sampling is stratified, selecting census sections. In the second stage, the sampling is systematic with a random start, selecting inhabited family dwellings where all members aged 16 and over are interviewed. Our target population is made up of the groups of employed men ($k = 1$) and employed women ($k = 2$). For the current analysis we do not include the autonomous cities of Ceuta and Melilla, so we have a total of D =50 provinces. Finally, we take data from the third (SLFS2022.3) and fourth (SLFS2022.4) quarters of 2022, since they are the latest published at the time of doing the statistical analysis.

Let s_{ktdr}, $s_{ktd} = \cup_{r=1}^{R} s_{ktdr}$, $s_{kt} = \cup_{d=1}^{D} s_{ktd}$, $s_k = \cup_{t=1}^{2} s_{kt}$ and $s = \cup_{k=1}^{2} s_k$ be subsets (samples) of U_{ktdr}, U_{ktd}, U_{kt}, U_k and U respectively. The direct estimators of Z_{ktdr}, N_{ktd} and $\overline{Z}_{ktdr}$ are

$$\hat{Z}_{ktdr}^{dir} = \sum_{j \in s_{ktd}} w_{ktdj} z_{ktdj,r}, \quad \hat{N}_{ktd}^{dir} = \sum_{j \in s_{ktd}} w_{ktdj}, \quad \hat{\overline{Z}}_{ktdr}^{dir} = \frac{\hat{Z}_{ktdr}^{dir}}{\hat{N}_{ktd}^{dir}},$$

where w_{ktdj} is the sample weight of individual $u_{ktdj} \in U_{ktd}$. The direct estimator of E_{td} is

$$\hat{E}_{td}^{dir} = \sum_{r=1}^{R} \hat{\overline{Z}}_{2tdr}^{dir} \frac{\hat{\overline{Z}}_{1tdr}^{dir}}{\frac{\hat{N}_{2td}^{dir}}{\hat{N}_{1td}^{dir}} \hat{\overline{Z}}_{2tdr}^{dir} + \hat{\overline{Z}}_{1tdr}^{dir}} = \frac{1}{\hat{N}_{2td}^{dir}} \sum_{r=1}^{R} \frac{\hat{Z}_{2tdr}^{dir} \hat{Z}_{1tdr}^{dir}}{\hat{Z}_{2tdr}^{dir} + \hat{Z}_{1tdr}^{dir}}, \quad t = 1, 2,\ d = 1, \ldots, D.$$

To overcome the lack of precision of the direct estimators $\hat{\overline{Z}}_{ktdr}^{dir}$, we may fit a MFH to the compositions $\hat{\overline{Z}}_{ktd}^{dir} = \big(\hat{\overline{Z}}_{ktd1}^{dir}, \ldots, \hat{\overline{Z}}_{ktdR-1}^{dir}\big)'$, $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$. However, such a MFH will not produce predictions of $\overline{Z}_{ktd}$ fulfilling the condition $\sum_{r=1}^{R} \overline{Z}_{ktdr} = 1$. Therefore, we apply the isometric logistic transformation of R-compositions onto $\mathbb{R}^{R-1}$ described by Egozcue and Pawlowsky-Glahn (2019). This is, a bijective mapping from the simplex

$$\mathcal{S}_e^{R-1} = \{(z_1, \ldots, _R) \in \mathbb{R}^R : z_1 > 0, \ldots, z_q > 0, z_1 + \ldots + z_R = 1\}.$$

onto $\mathbb{R}^{R-1}$ which is a standard Euclidean space that admits MFHs. Further, we use the simpler notation $z_{ktd} \triangleq \hat{\overline{Z}}_{ktd}^{dir}$, $z_{ktd} = (z_{ktd1}, \ldots, z_{ktdR-1})'$ for the direct estimators of compositions $\overline{Z}_{ktd}$, and $\sigma_{z,k_1t_1r_1k_2t_2r_2} = \mathrm{cov}_\pi(z_{k_1t_1r_1}, z_{k_2t_2r_2})$, $k_1, k_2 = 1, 2$, $t_1, t_2 = 1, 2$, $r_1, r_2 = 1, \ldots, R-1$, for the design-based variances and covariances. In matrix form, we have

$$\mathrm{var}_\pi(z_{ktd}) = \big(\sigma_{z,k_1t_1r_1k_2t_2r_2}\big)_{k_1,k_2=1,2;t_1,t_2=1,2,r_1,r_2=1,\ldots,R-1}.$$

For $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$, the ILR transformation of the composition z_{ktd} is $y_{ktd} = h(z_{ktd}) = (h_1(z_{ktd}), ..., h_{R-1}(z_{ktd}))'$, where

$$y_{ktdr} = h_r(z_{ktd}) = \sqrt{\frac{R-r}{R-r+1}} \log \frac{z_{ktdr}}{\Big(\prod_{j=r+1}^{R} z_{ktdj}\Big)^{1/(R-r)}},\ r = 1, \ldots, R-1$$

For ease of exposition, we often write $y_{ktd} = \text{ilr}(z_{ktd})$ and $z_{ktd} = \text{ilr}^{-1}(y_{ktd})$. Since the ILR transformation of z_{ktd} is applied, $\text{var}_\pi(z_{ktd})$ has to be transformed as well. For this, we use a Taylor approximation. The partial derivatives of y_{ktdi} with respect to z_{ktdj} are stated as

$$\begin{aligned}
\frac{\partial y_{ktdi}}{\partial z_{ktdj}} &= \frac{h_j(z_{ktd})}{\partial z_{ktdj}} = \sqrt{\frac{R-1}{R-i+1}}, \quad i,j = 1,\dots R-1, i = j \\
\frac{\partial y_{ktdi}}{\partial z_{ktdj}} &= \frac{h_j(z_{ktd})}{\partial z_{ktdj}} = -\sqrt{\frac{R-1}{R-i+1}}.\frac{1}{R-i}, \quad i,j = 1,\dots R-1, i > j \\
\frac{\partial y_{ktdi}}{\partial z_{ktdj}} &= \frac{h_j(z_{ktd})}{\partial z_{ktdj}} = 0, \quad i,j = 1,\dots R-1, i < j
\end{aligned} \tag{1}$$

These are collected in the Jacobian matrix at z_{ktd} with elements $H_{ij}(z_{ktd}) = \partial h_i(z_{ktd})/\partial z_{ktdj}$. At $z_0 = R^{-1}1_{R-1}$, the transformed value is $y_0 = h(z_0) = \text{ilr}(z_0) = 0_{R-1}$. A Taylor series expansion of the ILR transformation $\text{ilr}(z_{ktd})$ around z_0 yields to

$$y_{ktd} = \text{ilr}(z_{ktd}) \approx \text{ilr}(z_0) + H(z_0)(z_{ktd} - z_0), \tag{2}$$

where $H(z_0)$ is the Jacobian matrix given by (1). From (2), we get the approximated covariance matrix

$$\text{var}_\pi(y_{ktd}) \approx H_0 \text{var}_\pi(z_{ktd}) H_0'. \tag{3}$$

For estimating $\overline{Z}_{ktdr}$, $k = 1,2$,$t = 1,2$,$d = 1,\dots,D$, $r = 1,\dots,R-1$,this paper proposes to fit a MFH model to $y_{ktd} \overset{ind}{\sim} N_{R-1}(\mu_{ktd}, V_{ktd})$, where μ_{ktd} is a mean vector depending on unknown regression parameters and auxiliary variables and V_{ktd} is a covariance depending of some unknown parameters. Section 3 gives a flexible MFH model for y_{ktd}, $k = 1,2$, $t = 1,2$, $d = 1,\dots,D$, which allows positive and negative covariances.

3. THE MODEL AND THE PREDICTORS

Let $y_{ktd} = (y_{ktd1},\dots,y_{ktdR-1})'$ be the ilr transformation of $z_{ktd} = (z_{ktd1},\dots,z_{ktdR-1})'$. Let $\mu_{ktd} = E_\pi(y_{ktd}) = (\mu_{ktd1},\dots,\mu_{ktdR-1})'$ be the vector of design-based expectations of y_{ktd}. The compositional MFH model is defined in two stages. The sampling model is

$$y_{ktd} = \mu_{ktd} + e_{ktd}, \quad k = 1,2,\ t = 1,2\ d = 1,\dots,D, \tag{4}$$

where the vectors of random errors $e_{ktd} \sim N(0, V_{ektd})$ are independent and the $(R-1)\times(R-1)$ covariance matrices V_{ektd} are known. Following (3), we take $V_{ektd} = H_0 \hat{\text{var}}_\pi(z_{ktd}) H_0'$, where $\hat{\text{var}}_\pi(z_{ktd})$ is a direct estimator of the design-based variance $\text{var}_\pi(z_{ktd})$. We assume that μ_{ktdr} is linearly related to the row vector of explanatory variables $x_{ktdr} = (x_{ktdr1},\dots,x_{ktdrm_{kr}})$ and we define $X_{ktd} = \text{diag}\,(x_{ktd1},\dots,x_{ktdR})_{(R-1)\times m_k}$, where $m_k = \sum_{r=1}^{R-1} m_{kr}$. Let β_{kr} be a column vector of size m_{kr} containing the regression parameters for μ_{ktdr} and let $\beta_k = (\beta_{k1}',\dots,\beta_{kr}')'_{m_k\times 1}$. The linking model is

$$\mu_{ktd} = X_{ktd}\beta_k + u_{ktd}, \quad k = 1,2,\ t = 1,2,\ d = 1,\dots,D, \tag{5}$$

where the vectors of random effects u_{ktd}'s are independent of the vectors e_{ktd}'s, and

$$u_{ktd} \sim N(0, V_{uktd}),\ V_{uktd} = \underset{1\le r\le R-1}{\text{diag}}(\sigma_{kr}^2),\ k = 1,2,\ t = 1,2,\ d = 1,\dots,D.$$

Let $\theta = \underset{1\le k\le 2}{\text{col}}(\underset{1\le r\le R-1}{\text{col}}(\sigma_{kr}^2))$ be the column vector of variance parameters. Let I_n be the $n\times n$ identity matrix. At the level of U_{kt}, we define the following vectors and matrices

$$y_{kt} = \underset{1\le d\le D}{\text{col}}(y_{ktd}),\ u_{kt} = \underset{1\le d\le D}{\text{col}}(u_{ktd}),\ e_{kt} = \underset{1\le d\le D}{\text{col}}(e_{ktd}),$$

$$X_{kt} = \underset{1\le d\le D}{\text{col}}(X_{ktd}),\ Z_{kt} = I_{D(R-1)},\ V_{ukt} = \underset{1\le d\le D}{\text{diag}}\,(V_{uktd}),\ V_{ekt} = \underset{1\le d\le D}{\text{diag}}\,(V_{ektd}),$$

where col is the matrix operators stacking by columns. At the level of U, we define $m = \sum_{k=1}^{K} m_k$ and the following vectors and matrices

$$y = \underset{1\le k\le 2}{\text{col}}\left(\underset{1\le t\le 2}{\text{col}}(y_{kt})\right),\ u = \underset{1\le k\le 2}{\text{col}}\left(\underset{1\le t\le 2}{\text{col}}(u_{kt})\right),\ e = \underset{1\le k\le 2}{\text{col}}\left(\underset{1\le t\le 2}{\text{col}}(e_{kt})\right),\ \beta = \underset{1\le k\le 2}{\text{col}}(\beta_k),$$

$$X = \underset{1\le k\le 2}{\text{diag}}\left(\underset{1\le t\le 2}{\text{col}}(X_{kt})\right),\ Z = I_{KTDq},\ V_u = \underset{1\le k\le 2}{\text{diag}}\left(\underset{1\le t\le 2}{\text{diag}}\,(V_{ukt})\right),\ V_e = \underset{1\le k\le 2}{\text{diag}}\left(\underset{1\le t\le 2}{\text{diag}}\,(V_{ekt})\right).$$

In matrix form, the MFH model (4)-(5) is

$$y = X\beta + Zu + e, \tag{6}$$

where e, u are independent with distributions $e \sim N(0, V_e)$ and $u \sim N(0, V_u)$.

The components β_{kr} of the vector β in (6) depend on k and r, but not on d or t. By changing diag to col in the definition of X_{kd}, or in the definition of X, we can obtain regression coefficients depending only on k, only on r, or not depending on k and r. Since these variants can be written in the form of (6), we have developed R software for them but we do not treat them as different models. In what follows, we continue with the description of the model.

Under model (6), it holds that

$$E(y) = X\beta \quad \text{and} \quad V = \text{var}(y) = Z'V_uZ + V_e = V_u + V_e \triangleq \underset{1\le k\le 2}{\text{diag}}\big(\underset{1\le t\le 2}{\text{diag}}\big(\underset{1\le d\le D}{\text{diag}}\,(V_{ktd})\big)\big),$$

where $V_{ktd} = V_{uktd} + V_{ektd}$, $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$. Further, the best linear unbiased estimator (BLUE) of β, and the best linear unbiased predictors (BLUP) of u and μ are

$$\hat\beta^{blue} = (X'V^{-1}X)^{-1}X'V^{-1}y,\ \hat u^{blup} = V_uZ'V^{-1}(y - X\hat\beta^{blue}),\ \hat\mu^{blup} = X\hat\beta^{blue} + Z\hat u^{blup}. \tag{7}$$

The residual maximum likelihood (REML) log-likelihood of model (6) is

$$l_{reml}(\theta) = -\frac{2DTR - m}{2}\log 2\pi + \frac{1}{2}\log|X'X| - \frac{1}{2}\log|V| - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}y'Py,$$

where $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$, $PVP = P$ and $PX = 0$. For calculating $Q = X'V^{-1}X$, we may apply the alternative expression

$$Q = \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} x'_{ktd}V_{ktd}^{-1}x_{ktd}.$$

By taking partial derivatives of l_{reml} with respect to σ^2_{kr}, $k = 1, 2$, $r = 1, \ldots, R-1$, we obtain the score vector

$$S(\theta) = \underset{1\le k\le 2}{\text{col}}\big(\underset{1\le r\le R}{\text{col}}(S_{kr})\big),\ S_{kr} = S_{kr}(\theta) = \frac{\partial l_{reml}}{\partial\sigma^2_{kr}} = -\frac{1}{2}\text{tr}(PV_{kr}) + \frac{1}{2}y'PV_{kr}Py,\ V_{kr} = \frac{\partial V}{\partial\sigma^2_{kr}}.$$

Let δ_{ab} be the Kronecker delta. To calculate V_{kr}, we recall that

$$V_u = \underset{1\le k\le 2}{\text{diag}}\big(\underset{1\le t\le 2}{\text{diag}}\big(\underset{1\le d\le D}{\text{diag}}\big(\underset{1\le r\le R}{\text{diag}}\,(\sigma^2_{kr})\big)\big)\big),$$

so that

$$V_{kr} = \frac{\partial V_u}{\partial\sigma^2_{kr}} = \underset{1\le a\le 2}{\text{diag}}\big(\underset{1\le t\le 2}{\text{diag}}\big(\underset{1\le d\le D}{\text{diag}}\big(\underset{1\le b\le R}{\text{diag}}\,(\delta_{ak}\delta_{br})\big)\big)\big) = \underset{1\le a\le 2}{\text{diag}}\,(\delta_{ak}\big(\underset{1\le t\le 2}{\text{diag}}\big(\underset{1\le d\le D}{\text{diag}}\big(\underset{1\le b\le R}{\text{diag}}\,(\delta_{br})\big)\big)\big).$$

By taking again partial derivatives, changing the sign and taking expectations, we get the Fisher information matrix

$$F(\theta) = \left(F_{k_1r_1,k_2r_2}\right)_{k_1,k_2,=1,2;\,r_1,r_2,=1,\ldots,R-1},$$

where

$$F_{k_1r_1,k_2r_2} = F_{k_1r_1,k_2r_2}(\theta) = \frac{1}{2}\text{tr}(PV_{k_1r_1}PV_{k_2r_2}), \quad k_1, k_2 = 1, 2,\ r_1, r_2 = 1, \ldots, R-1.$$

The REML updating equation of the Fisher-scoring algorithm is

$$\theta^{k+1} = \theta^k + F^{-1}(\theta^k)S(\theta^k). \tag{8}$$

To obtain starting values, we separately fit a multivariate Fay-Herriot (MFH) model to the data set of U_k, $k = 1, 2$, by using the R package msae. For $k = 1, 2$, $r = 1, \ldots, R-1$, the starting values $\hat{\sigma}^2_{kr,0}$, are the REML estimators of $\sigma^2_{k1}, \ldots, \sigma^2_{kq}$ in the k-th MFH model.

The output of algorithm (8), $\hat{\theta}$, is the REML estimator of θ. By plugging $\hat{\theta}$ in V_u, we get $\hat{V}_u = V_u(\hat{\theta})$ and $\hat{V} = \hat{V}_u + V_e$. By substituting $\hat{V}_u$ in (7), we obtain the REML-EBLUP of $\mu = X\beta + Zu$, i.e.

$$\hat{\mu}^{eblup} = X\hat{\beta}^{eblue} + Z\hat{u}^{eblup},\ \hat{\beta}^{eblue} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y,\ \hat{u}^{eblup} = \hat{V}_u Z'\hat{V}^{-1}(y - X\hat{\beta}^{eblue}). \tag{9}$$

We denote the predictors of μ and u as $\hat{\mu} = \hat{\mu}^{eblup}$ and $\hat{u} = \hat{u}^{eblup}$, respectively. The REML estimator of β is $\hat{\beta} = \hat{\beta}^{eblue}$. The asymptotic distributions of the REML estimators $\hat{\theta}$ and $\hat{\beta}$,

$$\hat{\theta} \sim N_{2(R-1)}(\theta, F^{-1}(\theta)), \quad \hat{\beta} \sim N_m(\beta, (X'V^{-1}X)^{-1}),$$

can be used to construct $(1-\alpha)$-level confidence intervals for σ^2_{kr} and β_j, i.e.

$$\hat{\sigma}^2_{kr} \pm z_{\alpha/2}\, \nu^{1/2}_{kr,kr},\ k = 1, 2,\ r = 1, \ldots, R-1, \quad \hat{\beta}_j \pm z_{\alpha/2}\, q^{1/2}_{jj},\ j = 1, \ldots, m, \tag{10}$$

where $F^{-1}(\hat{\theta}) = (\nu_{k_1r_1,k_2r_2})_{k_1,k_2,=1,2;\, r_1,r_2,=1,\ldots,R-1}$, $(X'V^{-1}(\hat{\theta})X)^{-1} = (\nu_{ij})_{i,j=1,\ldots,m}$ and z_α is the α-quantile of the $N(0,1)$ distribution. For $\hat{\beta}_j = \beta_0$, the p-value for testing the hypothesis $H_0 : \beta_j = 0$ is

$$p\text{-value} = 2P_{H_0}(\hat{\beta}_j > |\beta_0|) = 2P(N(0,1) > |\beta_0|/\sqrt{\nu_{jj}}).$$

To introduce model-based predictors of proportions and EIs, we follow Krause et al.(2022) and we introduce predictors of the ilr transformation of $\mu_{ktdr} = E[y_{ktdr}|u_{ktdr}]$ under the MFH model, i.e.

$$\pi_{ktdr} = \text{ilr}_r^{-1}(\mu_{ktdr}) = \text{ilr}_r^{-1}(E[y_{ktd}|u_{ktd}]),\ r = 1, \ldots, R-1,$$

We propose the plug-in predictors, $\hat{\pi}^{in}_{ktd} = (\hat{\pi}^{in}_{ktd}, \hat{\pi}^{in}_{ktdR-1})'$ and $\hat{\pi}^{in}_{ktdR}$, of the proportions $\pi_{ktd} = (\pi_{ktd1}, \ldots, \pi_{ktdR-1})'$ and π_{ktdR} are jointly obtained by $\hat{\pi}_{ktd} = \text{ilr}^{-1}(\hat{\mu}^{eblup}_{ktd})$ or equivalently

$$\hat{\pi}_{ktdr} = \frac{\exp\{\hat{z}_{ktdr}\}}{\sum_{i=1}^{R-1}\exp\{\hat{z}_{ktdi}\}},\ \hat{z}_{ktd} = \Psi'\hat{\mu}^{eblup}_{ktd},\ r = 1, \ldots, R-1$$

where Ψ is a $(R-1) \times R$ transformation matrix defined by

$$\psi_{ij} = \frac{1}{\sqrt{(q-i)(q-i+1)}} \text{ for } i + j \leq q,\ \ \psi_{ij} = -\frac{\sqrt{q-i}}{\sqrt{q-i+1}} \text{ for } i + j = q + 1,$$

and 0 otherwise. The ilr transformation, $\hat{\mu}^{eblup}_{ktd}$ has been calculated in (9) and $\hat{\pi}_{ktdR} = 1 - \sum_{r=1}^{R-1}\hat{\pi}_{ktdr}$.

The plug-in predictors of the domain proportions $\overline{Z}_{ktdr}$ are $\hat{\overline{Z}}^{in}_{ktdr} = \hat{\pi}^{in}_{ktdr}$ and the plug-in predictors of the counts $Z_{ktdr} = N_{ktd}\overline{Z}_{kdr}$ are $\hat{Z}^{in}_{ktdr} = N_{ktd}\hat{\overline{Z}}^{in}_{ktdr}$, $r = 1, \ldots, R$. For every domain d, the predictors based on compositional models fulfill the two conditions: (1) $0 \leq \pi_{ktdr} \leq 1$, $r = 1, \ldots, R$, and (2) $\sum_{r=1}^{R}\pi_{ktdr} = 1$. Estimating π_{ktdr} with predictors based on univariate area-level mixed models, like Fay-Herriot, is not a good option because the conditions (1) and (2) might not be fulfilled.

The plug-in predictors of EIs are

$$\hat{E}^{in}_{td} = \sum_{r=1}^{R}\hat{\pi}^{in}_{2tdr}\frac{\hat{\pi}^{in}_{1tdr}}{\frac{\bar{N}^{dir}_{2td}}{\bar{N}^{dir}_{1td}}\hat{\pi}^{in}_{2tdr} + \hat{\pi}^{in}_{1tdr}}$$

To estimate the MSEs of the plug-in predictors, we follow a parametric bootstrap approach.

4. SIMULATION EXPERIMENTS

This section presents two simulation experiments that mimic the application to SLFS2022.3 and SFLS2022.4 data. We take the same auxiliary data and we generate the target vectors from the MFH model selected in Section 5. This to say, we assume the estimates of the model parameters in the application to SLFS2022 data are the true parameters. Simulation 1 investigates the performance of the fitting algorithm that calculates the REML estimators of the model parameters and studies the behaviour of the EI predictors. Simulation 2 deals with the MSE bootstrap estimation and provides a recommendation on the number of replicates to be used.

4.1 SIMULATION 1

The target of Simulation 1 is to investigate the behaviour of the fitting algorithm and the performance of the direct and plug-in predictors of E_{td}, $t = 1, 2$, $d = 1, \ldots, D$. We remind that there are $R = 7$ occupation sectors and $D = 50$ provinces. We run Simulation 1 with $I = 10^3$ iterations.

Table 1 presents the biases (BIAS) and the root-MSEs (RMSE) of the REML estimators $\hat{\beta}$ and $\hat{\theta}$ of the regression and variance parameters of the MFH model tha generates the data. For $k = 1, 2$, $r = 1, \ldots, 6$, the regression parameters are denoted as follows. The intercept of sex k and occupation sector r is β^0_{kr}. The slope of an auxiliary variable x for sex k and occupation sector r is β^x_{kr}. For $k = 1, 2$, $r = 1, \ldots, 6$, the variance parameters are $\theta_{kr} = \sigma^2_{kr}$. Table 1 shows that the biases and root-MSEs of the REML estimators of β and θ are very close to zero.

r	β	True	BIAS	RMSE	θ	True	BIAS	RMSE
1	β^0_{11}	-1.5657	-0.0016	0.3011	θ_{11}	0.0556	0.0948	0.1365
	β^0_{21}	1.6500	-0.0000	0.2550	θ_{21}	0.2458	-0.0952	0.1370
	β^{edu3}_{11}	0.6098	0.0056	0.4803				
	β^{age3}_{21}	-3.3970	0.0058	1.2027				
2	β^0_{12}	-3.1287	0.0000	0.2165	θ_{12}	0.0665	-0.0014	0.0194
	β^0_{22}	-1.4794	0.0111	0.3445	θ_{22}	0.0499	0.0113	0.0226
	β^{edu3}_{12}	4.2929	0.0015	0.3452				
	β^{situ4}_{22}	1.8401	-0.0181	0.5464				
3	β^0_{13}	-2.2628	0.0064	0.1972	θ_{13}	0.0573	0.0002	0.0178
	β^0_{23}	-3.2895	0.0032	0.3590	θ_{23}	0.0335	0.0107	0.0198
	β^{edu3}_{13}	2.6498	-0.0106	0.3147				
	β^{edu3}_{23}	2.9475	-0.0063	0.4799				
4	β^0_{14}	1.0644	-0.0023	0.2026	θ_{14}	0.0630	-0.0060	0.0145
	β^0_{24}	-1.8308	0.0017	0.4811	θ_{24}	0.0382	0.0095	0.0162
	β^{age3}_{14}	-2.8231	0.0095	0.9124				
	β^{age2}_{24}	2.3840	-0.0033	0.8615				
5	β^0_{15}	-0.0449	0.0074	0.2515	θ_{15}	0.0217	0.0151	0.0220
	β^0_{25}	0.7176	0.0046	0.2791	θ_{25}	0.0475	-0.0115	0.0192
	β^{nac1}_{15}	-0.4641	-0.0079	0.2906				
	β^{nac1}_{25}	-0.9795	-0.0047	0.3259				
6	β^0_{16}	0.2702	0.0365	0.3941	θ_{16}	0.0506	0.0450	0.0657
	β^0_{26}	0.9602	-0.0004	0.3928	θ_{26}	0.1353	-0.0414	0.0631
	β^{nac1}_{16}	-1.2658	-0.0415	0.4564				
	β^{nac1}_{26}	-2.9924	-0.0023	0.4557				

Table 1: Biases and root-MSEs of estimators of model parameters.

Table 2 presents the average absolute biases (ABIAS), the average absolute relative biases in % (ARBIAS),the average root-MSEs (RMSE) and the average relative root-MSEs in % (RRMSE) of the predictors of EIs. The average relative absolute biases and average relative root-MSEs are quite small in all cases, lower than 3%. We observe that for all these measures the plug-in predictor performs better than the direct estimator.

	$ABIAS$	$RMSE$	$ARBIAS$	$RRMSE$
$\hat{E}_2^{in}$	0.0017	0.0086	0.3965	1.9626
$\hat{E}_2^{dir}$	0.0027	0.0104	0.6347	2.3750

Table 2: Biases and root-MSEs of predictors.

4.2 SIMULATION 2

Simulation 2 studies the behavior of the estimator of the MSE of the predictors $mse^*(\hat{E}_{td}^{in})$ and $mse^*(\hat{E}_{td}^{dir})$. The target is to give a recommendation on the number of bootstrap replicates B to implement. To do this, such estimators are compared with the empirical MSE of $\hat{E}_{td}$ obtained at the output of Simulation 1.

Table 3 presents the average absolute biases AB_2 and the average root-MSEs RE_2 of the bootstrap estimators $mse^*(\hat{E}_{2d}^{in})$. Figure 1 presents the boxplots of the relative biases RB_{2d} (left) and the relative root-MSEs RRE_{2d} (right) of the bootstrap estimators $mse^*(\hat{E}_{td}^{in})$ of the EIs. The performance measures of the bootstrap estimators show a significant reduction in RRMSE from $B = 400$ onwards, although it does not stabilise. We therefore recommend $B = 400$, which improves on previous results and we believe is a relative root-mean-square error sufficiently accurate to be satisfied.

	$B = 50$	$B = 100$	$B = 200$	$B = 300$	$B = 400$	$B = 600$
AB_2^{dir}	1.8385e-05	1.8109e-05	6.3925e-06	6.5615e-06	6.49715e-06	6.5037e-06
AB_2^{in}	1.4074e-05	1.4427e-05	9.4536e-06	1.0868e-05	9.3951e-06	9.3961e-06
RE_2^{dir}	3.2508e-05	2.6817e-05	1.9954e-05	1.6212e-05	1.5981e-05	1.4280e-05
RE_2^{in}	2.7959e-05	2.3067e-05	1.6162e-05	1.5958e-05	1.4081e-05	1.3250e-05

Table 3: AB_2 and RE_2 of mse_2^{*in} for $B = 50, 100, 200, 300, 400, 600$.

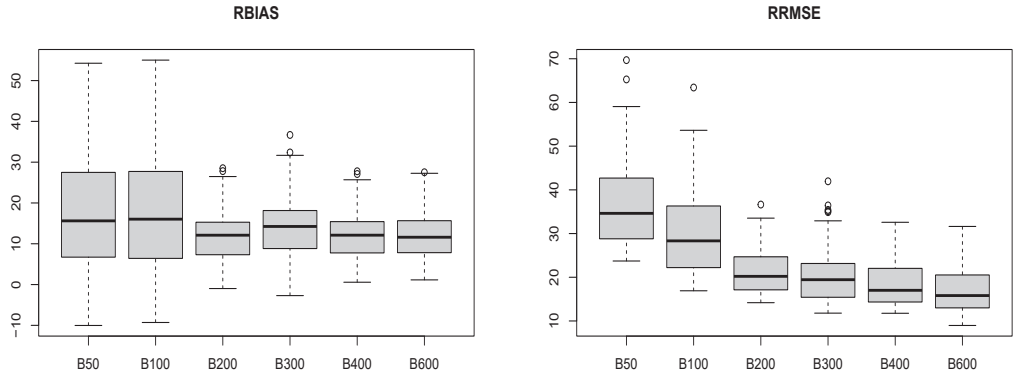


Figure 1: $RBIAS_{2d}$ (left) and RRE_{2d} (right) of mse_{2d}^{*in} for $B = 50, 100, 200, 300, 400, 600$.

5. APPLICATION TO REAL DATA

This section applies the introduced statistical methodology to the SLFS2022.3 and SLFS2022.4 data described in Section 2. To overcome the lack of precision of the direct estimators, we construct a data file with auxiliary aggregated variables. The selected auxiliary variables are direct estimates of the means (proportions) of unit-level binary variables, measured on individuals of U_{ktdr}. They are.

Age group, with 3 categories: between 16 and 30 years (age1), between 30 and 50 years (age2) and over 50 years (age3).

Citizenship, with 2 categories: Spanish (nac1) and not Spanish (nac2).
Education, with 3 categories: primary or less (edu1), basic secondary education (edu2), bachelor or higher education, such as university (edu3):
Working hours, with 2 categories: full-time (work1) and part-time work (work2):
Professional status, with 5 categories: employer (with or without employees) or independent worker (st1), cooperative or family business (st2), public sector salaried employee (st3), private sector salaried employee (st4) and others (st5).

We apply the ilr transformation to the target data, i.e. $y_{ktd} = \text{ilr}(z_{ktd})$, $k = 1, 2$, $t = 1, 2$, $d = 1, \ldots, D$. The MFH model is fitted to the transformed data and to a subset of significant auxiliary aggregated variables. For every province, time period and sex, the preliminary population indicators of interest are the proportions of employed people in the seven occupation sectors. We exclude the autonomous cities of Ceuta and Melilla from the statistical analysis. We are interested in estimating domain compositions with $R = 7$ categories. As explanatory variables we select those with a consistent actual interpretation and better in terms of p-value when fitting the MFH model to the data. Table 4 presents the selected auxiliary variables, the estimates of the regression parameters, the standard errors (SE) and the p-values. Since all the auxiliary variables are proportions, the sign and magnitude of the regression parameters give interesting interpretations under the assumption that the rest of the auxiliary variables remain constant.

r	k	β	Estimate	SE	p-value	β	Estimate	SE	p-value
1	1	β^0_{11}	-1.5657	0.1742	0.0000	β^{edu3}_{11}	0.6098	0.2769	0.0276
	2	β^0_{21}	1.6500	0.3409	0.0000	β^{age3}_{21}	-3.3970	1.6017	0.0339
2	1	β^0_{12}	-3.1287	0.2174	0.0000	β^{edu3}_{12}	4.2929	0.3423	0.0000
	2	β^0_{22}	-1.4794	0.3084	0.0000	β^{st4}_{22}	1.8401	0.4851	0.0001
3	1	β^0_{13}	-2.2628	0.2005	0.0000	β^{edu3}_{13}	2.6498	0.3171	0.0000
	2	β^0_{23}	-3.2895	0.3489	0.0000	β^{edu3}_{23}	2.9475	0.4621	0.0000
4	1	β^0_{14}	1.0644	0.2191	0.0000	β^{age3}_{14}	-2.8231	0.9874	0.0042
	2	β^0_{24}	-1.8308	0.4533	0.0001	β^{age2}_{24}	2.3840	0.8118	0.0033
5	1	β^0_{15}	-0.0449	0.2117	0.8317	β^{nac1}_{15}	-0.4641	0.2457	0.0589
	2	β^0_{25}	0.7176	0.3030	0.0178	β^{nac1}_{25}	-0.9795	0.3530	0.0055
6	1	β^0_{16}	0.2702	0.2847	0.3426	β^{nac1}_{16}	-1.2658	0.3297	0.0001
	2	β^0_{26}	0.9602	0.4747	0.0431	β^{nac1}_{26}	-2.9924	0.5522	0.0000

Table 4: Estimates, standard errors and p-values of regression parameters.

Table 5 gives the asymptotic 95% confidence intervals (CI), introduced in (10), of the variance parameters $\theta_{kr} = \sigma^2_{ukr}$, $k = 1, 2$, $r = 1, \ldots, 6$. This table shows that all variances are significantly greater than zero.

	k	θ_{k1}	θ_{k2}	θ_{k3}	θ_{k4}	θ_{k5}	θ_{k6}
$\hat{\theta}$	1	0.0556	0.0664	0.0572	0.0629	0.0216	0.0506
CI inf		0.0398	0.0426	0.0364	0.0432	0.0121	0.0339
CI sup		0.0714	0.0902	0.0780	0.0826	0.0311	0.0673
$\hat{\theta}$	2	0.2457	0.0499	0.0334	0.0381	0.0474	0.1352
CI inf		0.1769	0.0293	0.0154	0.0227	0.0304	0.0939
CI sup		0.3146	0.0704	0.0515	0.0535	0.0644	0.1765

Table 5: Asymptotic 95% CIs for θ_{kr}, $k = 1, 2$, $r = 1, \ldots, 6$, .

Figure 2 shows the boxplots of the standardized residuals (sresiduals) from the MFH model for men (left) and women (right) by occupation sector. The sresiduals are quite symmetric around zero and do not present any relevant pattern. The model-based predictions of the proportions of employed men seem to be greater than the corresponding direct estimates in Sector 7. As before, it can also be seen that the residuals are symmetric around zero, but in this case it can be noted that there exist some dispersion. Further, there are few sresiduals (33 among 1400) outside the

interval $(-3, 3)$ so we consider that outliers do not play a relevant role in the performance of the EBLUPs. Figures 2 illustrate that the residual diagnostics do not present large deviation from normality.

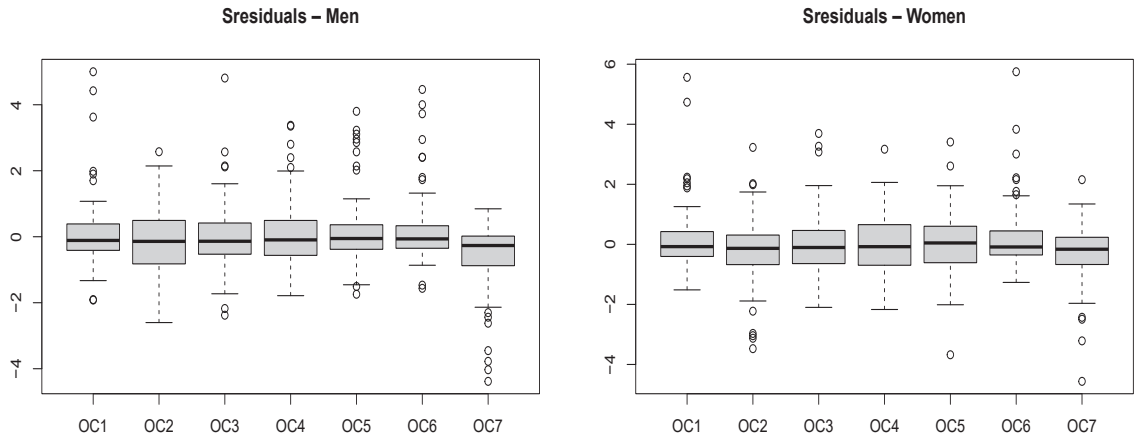


Figure 2: Boxplot of standardized residuals for men (left) and women (right) by occupation sector.

Figure 3 (left) plots the direct and compositional plug-in estimated EIs at fourth quarter for provinces sorted by sample size. There are strong similarities between EIs estimates for plug-in predictors and direct estimators. We can appreciate that most of estimates are between 0.4 and 0.5, which indicates small variation for the index. This leads us to think that there is homogeneity in gender segregation for the Spanish provinces. Figure 3 (right) plots the bootstrap MSE estimates of $\hat{E}_{2d}^{in}$ and $\hat{E}_{2d}^{dir}$, sorted by province-time sample size $n_{.2d} = n_{12d} + n_{22d}$, $d = 1, \ldots, D$. This figure shows that the estimated MSEs of plug-in predictor are lower than the corresponding ones of the direct estimator. Although it seems to be large sample sizes, in fact, we have to take into account that each predictor component of the sum is obtained by sex-province-sector cross, so the sample size is small.

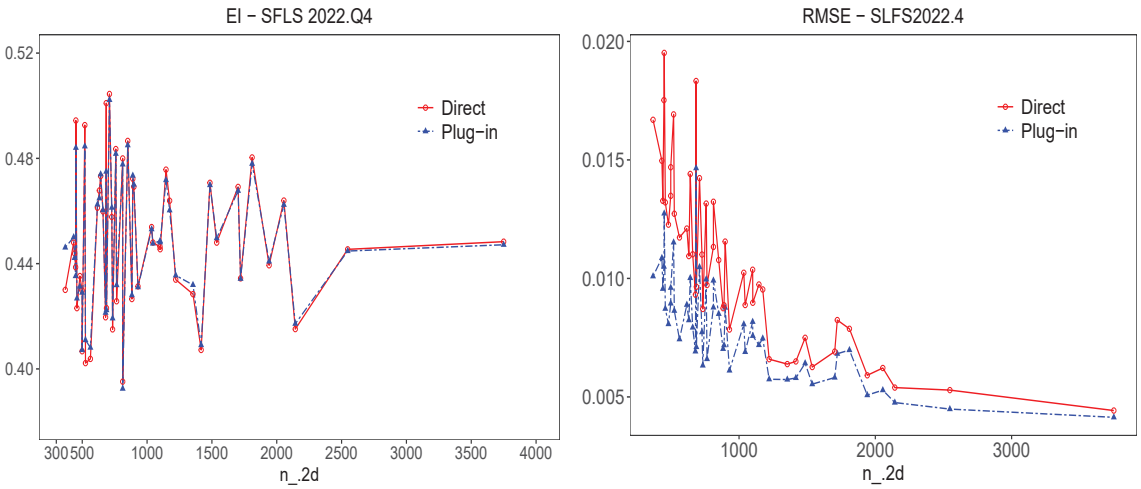


Figure 3: EI estimates (left) and RMSE estimates of direct and plug-in predictors of EIs (right), sorted by sample size for SFLS2022.Q4.

Figure 4 (left) presents a map of the Spanish provinces coloured according to the Exposure predictions for the fourth quarter of 2022. Figure 4 (right) gives the maps of their relative root-MSEs. These maps allow us to analyze segregation between sexes across Spanish provinces. Firstly, we

observe that the higher discrepancies in the distribution of employed men and women by main occupations are mostly found in the north of the peninsula: Galicia, Catalonia, Castile-Leon, Asturias, Cantabria, Navarra, País Vasco, La Rioja, Aragon and Madrid. This leads us to believe that seems to be a pattern of gender exposure in this geographical area. Thirdly, the south zone (Andalusia, Extremadura, Castile-La Mancha, Valencian Community and Canary Islands) appears to have medium-high gender exposure. Finally, the men and women distributions in provinces with similar demographic and socioeconomic conditions have, in general, similar exposure.

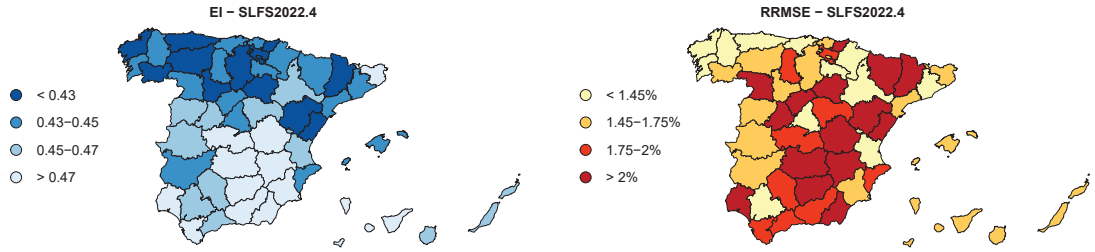


Figure 4: EI estimates and RRMSE in SFLS2022.4.

Figure 4 (left) shows that estimates of Exposures are not close to one (highest value is ≈ 0.516). In terms of labour equality, those low predicted values reveal the magnitude of the problem: the labour market disadvantages women and the occupational distribution is clearly non-homogeneous. According to our results, public and private institutions should implement measures of work equality and promote the inclusion of men and women in those sectors in which their presence is minority. Moreover, figure 4 (right) presents maps of the estimated relative root-MSEs of the exposure index plug-in predictors for third and fourth quarter respectively. The estimates of relative root-MSEs are greater for women than for men, although none of them exceeds 4%. In both cases the estimates are particularly low, this is because the variance is small for both time periods, which makes it possible to predict with high accuracy ($\hat{\text{var}}(E_{2d}) \in [0.0005, 0.0008]\, \forall d = 1, \ldots, 50$). The results obtained in Figure 4 were obtained by a parametric bootstrap approach with $B = 500$ iterations.

To understand where and how the differences between the men and women occupation sector distributions occur, we calculate for each occupational sector the average across provinces of their summands in the formula of plug-in predictors of EIs, i.e.

$$\hat{E}^{in}_{2r} = \frac{1}{D}\sum_{d=1}^{D} \hat{\pi}^{in}_{2dr} \frac{\hat{\pi}^{in}_{1dr}}{\frac{\hat{N}^{dir}_{22d}}{\hat{N}^{dir}_{12d}}\hat{\pi}^{in}_{2dr} + \hat{\pi}^{in}_{1dr}}, \quad r = 1, \ldots, 7,$$

and we estimate their corresponding root-MSEs, $rmse(\hat{E}^{in}_{2r})$, and coefficient of variation in %, $\text{cv}(\hat{E}^{in}_{2r}) = 100\, rmse(\hat{E}^{in}_{2r})/|\hat{E}^{in}_{2r}|$, by parametric bootstrap. Table 6 contains the sector contributions $\hat{E}^{in}_{2r}$, $r = 1, \ldots, 7$ and the corresponding estimates of their root-MSEs and of their coefficient of variations.

Contribution	OC1	OC2	OC3	OC4	OC5	OC6	OC7
$\hat{E}^{in}_{2r}$	0.0162	0.0930	0.0548	0.0439	0.1116	0.0848	0.0439
$rmse(\hat{E}^{in}_{2r})$	0.0001	0.0012	0.0006	0.0004	0.0015	0.0012	0.0003
$\text{cv}(\hat{E}^{in}_{2r})$	0.9431	1.3434	1.1618	1.0487	1.3291	1.4080	0.6905

Table 6: Occupational average EI values.

Table 6 gives contribution by each occupation sector of EI estimates. It shows that lowest exposure within occupational sectors are found in Sectors 1, 4 and 7 in terms of lower mean. This indicates that the segregation between men and women in these sectors is, generally, higher than the other sectors between the provinces. On the other hand, Sectors 2, 5 and 6 have greater mean

and CV values, showing that in those sectors, in general, the presence of men and women is more homogeneous in comparaison. This allows us to understand the results shown above: there is a clear majority of men employed in jobs related to management of companies and administrations (Sector 1), administrative and office employees (Sector 4) and industrial jobs such as mining, manufacturing and construction (Sector 7). This differences reduce the rate of exposure between men and women by province. The major presence of men in this sector is, we believe, due to historical and social reasons.

6. CONCLUSIONS

The EI serves as a significant statistical index utilized in sociological investigations to quantify the level of group segregation within various occupational sectors. In cases where the sample sizes are sufficiently large, direct methods can be employed to estimate the EIs. However, when aiming to estimate EIs at disaggregated levels with limited sample sizes, alternative model-based estimation procedures must be employed. This approach allows us to capitalize on the additional information integrated into the models, thus yielding more precise predictions.

When examining compositional data using multivariate linear mixed models, a flexible modeling strategy arises, as it can adapt to the hierarchical structure of the population based on factors such as gender, time period, province, and occupation sector. By employing a logratio transformation of the direct estimators of proportions by province and time period for both men and women within each labor sector, an MFH model can be proposed. This model can then incorporate auxiliary information to enhance the accuracy of predictions. In this context, plug-in predictors have been considered. In order to estimate mean squared errors (MSEs), a parametric bootstrap method has been applied and we advise to use at least $B = 400$ iterations, as a good spot between accuracy and computational time.

In the application of these methodologies to the SLFS2022.3 and SLFS2022.4 datasets, the introduced plug-in predictors have been utilized. Moreover, the study presents findings concerning the components of Exposure in particular, for OC1 (Directors, managers and management of companies and public administrations), OC4 (accounting, administrative and other office employees) and OC7 (plant and machinery operators and assemblers, artisans and skilled workers in the manufacturing, construction and mining industries categories). The plug-in predictors have exhibited lower MSE values in comparison to the direct Hájek estimators. By mapping the Exposure Indexes (EIs) in Spain, most of the predictions have displayed an estimated relative root mean squared error (RRMSE) below 10%, which signifies a satisfactory level of accuracy for small area estimation (SAE) problems. Ultimately, the study has identified the provinces that experience the greatest impact from gender occupational segregation. Finally, we conclude that using recent past data significantly improves the results of the study both in simulations and in application to real data.

ACKNOWLEDGEMENTS

This work has been carried out thanks to the financial support of the Project "Data Science and Sustainable Development Goals (DATOS)" (PROMETEO/2021/063 grant) co-financed by the State Program for the Generation of Knowledge and Scientific and Technological Strengthening of the R&D System (PGC2018-096840-B-I00 grant). The authors also thank the Operational Research Center of the Miguel Hernández University of Elche for the resources provided.

REFERENCES

Bell, W. (1954). A probability model for the measurement of ecological segregation. Social Forces, 32(4), 357-364.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. Test, 28(3), 599-638.

Massey, D. S., Denton, N. A. (1988). The dimensions of residential segregation. Social Forces, 67(2), 281-315.

Morales, D., Esteban, M.D., Pérez, A., Hobza, T. (2021). A course on Small Area Estimation and mixed models. Springer.

Pratesi, M. (2016). Analysis of poverty data by Small Area Estimation. Wiley Series in Survey Methodology, John Wiley.

Rao J.N.K., Molina, I. (2015). Small Area Estimation, 2nd Edition, Wiley, New York.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

DETERMINACIÓN DA IMPORTANCIA DE VARIABLES FÍSICAS DO TERRITORIO EN APROVEITAMENTOS FORESTAIS MEDIANTE ANÁLISE CONXUNTO

David Iglesias González[1], Laura Alonso Martínez[2], Juan Picos Martín[2], Mª Carmen Iglesias Pérez[3] e Julia Armesto González[2]

[1]Universidade de Vigo, Máster en Técnicas Estadísticas.
[2]Universidade de Vigo, Escola de Enxeñaría Forestal.
[3]Universidade de Vigo, Departamento de Estatística e Investigación Operativa.

RESUMO

O obxectivo deste traballo foi explorar e cuantificar as preferencias dos profesionais do sector forestal galego na elección dunha parcela para o seu aproveitamento forestal, co fin de proporcionar unha ferramenta que permita identificar a potencialidade dunha parcela para o aproveitamento en función das características físicas do terreo. Para coñecer as súas preferencias, realizamos unha análise conxunta baseada en eleccións, tomando como atributos do modelo cinco variables físicas comúns das parcelas forestais que poden chegar a condicionar a realización dun aproveitamento: a superficie da parcela, a forma da parcela, a fragmentación circundante, a distancia á estrada máis próxima e a pendente da parcela. Utilizando técnicas de deseño experimental, elaboramos unha enquisa e distribuímola entre os profesionais do sector. Sobre as respostas aplicamos unha regresión loxística multinomial e estimamos as utilidades parciais para cada unha das variables. A partir destas utilidades é posible construír un indicador que permita identificar a utilidade de calquera parcela en Galicia e representala nun mapa. Esta información pode facilitar unha xestión máis eficiente dos recursos forestais de cara a mellorar a competitividade da cadea de valor forestal de Galicia.

Palabras e frases chave: análise conxunto, utilidades parciais, xestión forestal, cadea de valor.

REFERENCIAS

Rao Vithala, R. (2014) *Applied Conjoint Analysis*. Springer.

Xunta de Galicia, Consellería do Medio Rural (2021). *1ª Revisión del plan forestal de Galicia, hacia la neutralidad carbónica*. Xunta de Galicia, C 1946-2021.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Polynomial Optimization: Tightening branch and bound schemes with conic constraints

Brais González-Rodríguez[1], Raúl Alvite-Pazó[2], Samuel Alvite-Pazó[2], Bissan Ghaddar[3], and Julio González-Díaz[1,2]

[1]Department of Statistics, Mathematical Analysis and Optimization and MODESTYA Research Group, University of Santiago de Compostela
[2]CITMAga (Galician Center for Mathematical Research and Technology)
[3]Ivey Business School, Western University, London, Ontario, Canada

ABSTRACT

Conic optimization has recently emerged as a powerful tool for designing tractable and guaranteed algorithms for non-convex polynomial optimization problems. On the one hand, tractability is crucial for efficiently solving large-scale problems and, on the other hand, strong bounds are needed to ensure high quality solutions. In this research, we investigate the strengthening of RLT relaxations of polynomial optimization problems through the addition of nine different types of constraints that are based on linear, second-order cone, and semidefinite programming to solve to optimality the instances of well established test sets of polynomial optimization problems. We describe how to design these conic constraints and their performance with respect to each other and with respect to the standard RLT relaxations. Our finding is that the different variants of nonlinear constraints (second-order cone and semidefinite) are the best performing ones in around 50% of the instances. Additionally, we present a machine learning approach to decide on the most suitable constraints to add to a given instance. The computational results show that the machine learning approach significantly outperforms each of the nine individual approaches.

Keywords: Global Optimization, Reformulation-Linearization Technique, Polynomial Programming, Conic Optimization, Machine Learning.

1. INTRODUCTION

The large volume of theoretical and computational research on conic optimization has led to important advances over the last few years in the efficiency and robustness of the associated algorithmic procedures to solve them. Leading state-of-the-art mixed integer linear programming (MILP) solvers such as `Gurobi` (Gurobi Optimization, 2022), `CPLEX` (IBM Corp., 2022), and `Xpress` (FICO, 2022) have recently added functionalities that allow to efficiently solve second-order cone programming (SOCP) problems and, further, `Mosek` (MOSEK ApS, 2022; Andersen and Andersen, 2000) has positioned itself as a reliable solver for general semidefinite programming (SDP) problems.

As a result of the convex nature of conic optimization problems, they are considered a powerful tool in designing branch-and-bound algorithms for general non-convex mixed integer nonlinear programming (MINLP) problems and, especially, polynomial optimization problems. The developments of the last several decades have shown that conic optimization is a central tool in addressing non-convexities. These advances have been more prominent in the case of polynomial optimization problems, and even more so in the particular case of quadratically-constrained quadratic optimization problems (QCQP), where a variety of convex relaxations have been thoroughly studied (Shor, 1987; Ghaddar et al., 2011; Burer and Ye, 2020; Bonami et al., 2019; Elloumi and Lambert, 2019). These conic-based relaxations include semidefinite programming and second-order cone programming as their powerhouses. They often provide fast and guaranteed approaches for computing bounds on the global value of the non-convex optimization problem at hand.

One of the primary goals of this work is to show that, nowadays, general branch-and-bound schemes can benefit from the inclusion of conic SOCP and SDP constraints. We develop our analysis for general polynomial optimization problems in the context of the Reformulation-Linearization Technique (RLT), for which it is common to tighten the linear relaxations using linear SDP-based cuts (Sherali et al., 2012; Baltean-Lugojan et al., 2019; González-Rodríguez et al., 2020), but the efficacy of adding SOCP or SDP constraints has not been studied. Arguably, the main reason why this last approach has received less attention so far is that the resulting algorithms need to rely on SDP solvers which, until recently, were probably not reliable enough and too expensive computationally in some instances.

In order to achieve the above goal, in this work we compare the performance of different linear SDP-based constraints as well as SOCP and SDP constraints. These conic constraints have to be carefully chosen since, for them to be competitive, it is important that the potential reduction in the size of the branch-and-bound tree coming from the additional tightness does not become overshadowed by the extra time required to solve each node. With this trade-off in mind, we consider constraints that are efficient in the sense of preserving the size and sparsity of the original RLT-based relaxation for polynomial optimization problems introduced in Sherali and Tuncbilek (1992) and further refined in Dalkiran and Sherali (2013). We consider a total of nine different versions of constraints to be added: one based on linear SDP-based cuts, four based on SOCP constraints, and four based on SDP constraints. These conic constraints are then integrated into the polynomial optimization solver `RAPOSa` (González-Rodríguez et al., 2020), whose core is an RLT-based branch-and-bound algorithm. As a second step, we then develop thorough computational studies on well established benchmarks and one of the main findings is that, in around 50% of the instances, the best performance is achieved by one of the versions explicitly incorporating SOCP or SDP conic constraints. The remaining 50% is split quite evenly between the baseline RLT and the version that incorporates SDP-based linear cuts. Importantly, the analysis also allowed identifying particular classes of problems where one specific family of SOCP/SDP conic constraints is consistently superior to the linear versions.

To the best of our knowledge, our contribution represents one of the very few implementations of branch-and-bound schemes with (non-linear) conic relaxations for broad classes of problems, with the added value of the generality of the resulting scheme, since it can be applied to any given polynomial optimization problem. The most related approaches are Burer and Vandenbussche (2008) for non-convex quadratic problems with linear constraints and Buchheim and Wiegele (2013) for unconstrained mixed-integer quadratic problems.

Last, but not least, in our computational experiments we also observe that there is a lot of variability in the best performing version of conic constraints for different instances, with each version beating the rest in a nonnegligible number of instances. This observation motivates the last contribution of this work, in which we exploit this variability by learning to choose the best version among our portfolio of different constraints. Building upon the framework in Ghaddar et al. (2022), we show that the resulting machine learning version significantly outperforms each and every one of the underlying versions. This last contribution naturally fits into the rapidly emerging strand of research on "learning to optimize", whose advances are nicely presented in the survey papers Lodi and Zarpellon (2017) and Bengio et al. (2021), and more recently in Kannan et al. (2022) for QCQP problems. The closest approach to this last part of our contribution is Baltean-Lugojan et al. (2019), in which deep neural networks are used to rank linear SDP-based cuts for quadratic problems. Then, only the top scoring cuts are added, aiming to obtain a good balance between the tightness and the complexity of the relaxations. A common feature of the linear SDP-based cuts used in Baltean-Lugojan et al. (2019) and those in this work is that they are designed with a big emphasis on sparsity considerations (number of nonzeros in the constraints), with the goal of obtaining computationally efficient relaxations. From the point of view of the learning process, the approaches are quite different since they focus on one type of constraint (the SDP-based cuts) and they want to learn the best SDP-based cuts to add at a given node, whereas we want to learn to choose for any given instance which conic constraints to include in the relaxations.

The contribution of this work can be summarized as follows. First, we study the potential of different (nonlinear) SOCP and SDP tightenings of the classic (linear) RLT relaxations for general polynomial optimization problems in a branch-and-bound scheme. Second, we design a machine

learning approach to learn the best tightening approach to use on a given instance and obtain promising results.

The remainder of this document is organized as follows. In Section 2 we present a brief overview of the classic RLT scheme. In Section 3 we describe the different families of conic constraints that will be integrated within the baseline RLT implementation. In Section 4 we present a first series of computational results. Then, in Section 5 we show how the conic constraints can be further exploited within a machine learning framework. Finally, we conclude in Section 6 and discuss future research directions.

2. FOUNDATIONS OF THE RLT TECHNIQUE

The Reformulation-Linearization Technique was originally developed in Sherali and Tuncbilek (1992). It was designed to find global optima in polynomial optimization problems of the following form:

$$\begin{array}{lll} \text{minimize} & \phi_0(\mathbf{x}) & \\ \text{subject to} & \phi_r(\mathbf{x}) \geq \beta_r, & r = 1, 2, \ldots, R_1 \\ & \phi_r(\mathbf{x}) = \beta_r, & r = R_1 + 1, \ldots, R \\ & \mathbf{x} \in \Omega \subset \mathbb{R}^n, & \end{array} \tag{PO}$$

where $N = \{1, \ldots, n\}$ denotes the set of variable indices, each $\phi_r(\mathbf{x})$ is a polynomial of degree $\delta_r \in \mathbb{N}$ and $\Omega = \{\mathbf{x} \in \mathbb{R}^n : 0 \leq l_j \leq x_j \leq u_j < \infty, \forall j \in N\} \subset \mathbb{R}^n$ is a hyperrectangle containing the feasible region. Then, $\delta = \max_{r \in \{0,\ldots,R\}} \delta_r$ is the degree of the problem and (N, δ) represents all possible monomials of degree δ.

The Reformulation-Linearization Technique consists of a branch and bound algorithm based on solving linear relaxations of the polynomial problem (**PO**). These linear relaxations are built by working on a lifted space, where each monomial of the original problem is replaced with a corresponding RLT variable. For example, associated to monomials of the form $x_1x_2x_4$ and $x_1^2x_3^2$ one would define the RLT variables X_{124} and X_{1133}, respectively. More generally, RLT variables are defined as

$$X_J = \prod_{j \in J} x_j, \tag{1}$$

where J is a multiset containing the information about the multiplicity of each variable in the underlying monomial. Then, at each node of the branch-and-bound tree, one would solve the corresponding linear relaxation. Whenever we get a solution of a linear relaxation in which the identities in (1) hold, we get a feasible solution of (**PO**). Otherwise, the violations of these identities are used to choose the branching variable, leading to the definition of two new nodes of the tree, with the corresponding linear relaxations.

In order to get tighter relaxations and ensure convergence, new constraints, called bound-factor constraints, must be added. They are of the following form:

$$F_\delta(J_1, J_2) = \prod_{j \in J_1} (x_j - l_j) \prod_{j \in J_2} (u_j - x_j) \geq 0.$$

Thus, for each pair of multisets J_1 and J_2 such that $J_1 \cup J_2 \subset (N, \delta)$ and $|J_1 \cup J_2| = \delta$, the corresponding bound-factor constraint is added to the linear relaxation. Dalkiran and Sherali (2013) show that it is not necessary to add all bound-factor constraints to the linear relaxations, since certain subsets of them are enough to ensure the convergence of the algorithm. More precisely, they proved that convergence to a global optimum of (**PO**) only requires the inclusion in the linear relaxation of those bound-factor constraints where $J_1 \cup J_2$ is a monomial that appears in (**PO**), regardless of its degree. Further, they also showed that convergence is also preserved if, whenever the bound-factor constraints associated with a monomial J are present, all bound-factor constraints associated with monomials $J' \subset J$ are removed. Motivated by these results, J-sets are defined as those monomials of degree greater than one present in (**PO**) which, moreover, are not included in any other monomial (multiset inclusion).

The analysis developed in this work builds upon the above theoretical results from Dalkiran and Sherali (2013) and, therefore, only the bound-factor constraints associated with J-sets are incorporated into the linear relaxations of (**PO**).

3. CONIC ENHANCEMENTS OF RLT

As already discussed earlier, the use of semidefinite programming to improve the performance of branch-and-bound schemes is not new, and Baltean-Lugojan et al. (2019) provides a thorough and up-to-date review of the field. Typically, the goal is to rely on semidefinite programming to tighten the relaxations of the original non-convex optimization problems targeted by a branch-and-bound algorithm. We start this section by reviewing the main ingredient of such methods and then present various families of SDP-driven constraints that can be incorporated into the RLT relaxations in an efficient way. In particular, they should preserve the underlying dimensionality and sparsity of the problem, which is crucial for these approaches to be competitive.

The main ingredient that has to be specified is the matrix or matrices on which positive semidefiniteness is to be imposed. To each (multi-)set of variables $\{x_j\}_{j\in J}$, with $J \subset (N, \delta)$, one can associate a vector $\omega = (x_j)_{j\in J}$ (resulting from the concatenation of all the variables in J, including repetitions). To any such vector one can associate matrix $M = \omega\omega^T$, which is trivially positive semidefinite. Now, let $M_L = [\omega\omega^T]_L$ be the matrix obtained when each monomial in M is replaced by the corresponding RLT variable in the lifted space. The constraint $M_L \succcurlyeq 0$ is a valid cut because it never removes feasible solutions of (**PO**) and, hence, it does not compromise convergence of the RLT algorithm to a global optimum of (**PO**). In practice, vectors of the form $(1, (x_j)_{j\in J})$ are often preferred, since they result in M_L matrices containing also variables in the original space and not only RLT variables, leading to tighter relaxations. In Sherali et al. (2012), for instance, the authors discuss different ways of defining w for a given J. Specifically, they mainly work with $\omega^1 = (x_j)_{j\in J}$, $\omega^2 = (1, (x_j)_{j\in J})$, and $\omega^3 = (1, x_1, x_2, \ldots)$, where ω^3 is defined by concatenating also monomials of degree greater than one, while ensuring that no monomial in the resulting matrix M has degree larger than δ, the degree of (**PO**).

We now move to the specification of the SDP-driven constraints for the RLT algorithm that constitute the subject of study in this work. As mentioned above, similar constraints have been discussed in the past, and the purpose of this section is to be explicit about how we have adapted past approaches to our setting.

3.1. LINEAR SDP-BASED CONSTRAINTS

In Sherali et al. (2012), the authors associate linear cuts to the constraints of the form $M_L \succcurlyeq 0$ as follows. At each node of the branch-and-bound tree, the positive semidefiniteness of the chosen M_L matrices is assessed at the solution of the corresponding relaxation. Given a negative eigenvalue of one such matrix with α as its associated eigenvector, then the valid cut $\alpha^T M_L \alpha \geq 0$ can be added to the linear relaxation to separate the current solution. A potential drawback of these cuts is that they may be very "dense", in the sense of involving a large number of variables, which may increase the solving time of the relaxations. Thus, as already discussed in Sherali et al. (2012), it is important to carefully choose M_L matrices.

The sparsity of the cuts is particularly important if the RLT algorithm is being run with the J-set approach since, in general, the resulting cuts might involve monomials not contained in any J-set. This would require to include additional RLT variables in the relaxations (and the corresponding bound-factor constraints), increasing the size and potentially reducing the sparsity of the relaxations. Here we follow González-Rodríguez et al. (2020), where the authors consider, at each node, all M_L matrices obtained from vectors ω^k associated with the different J-sets of (**PO**), which lead to sparse cuts that essentially preserve the dimensionality of the resulting relaxations. The authors present a detailed computational analysis, comparing different approaches to add inherit cuts. We adopt the best performing version, which consists of using vector ω^2 and inheriting all cuts from one node to all its descendants.

3.2. SDP CONSTRAINTS

We next describe two approaches to tighten the classic RLT relaxations by directly adding semidefinite constraints. They just differ in the matrices on which positive semidefiniteness is imposed.

Approach 1. For each J-set J, ω^1 is used to define the M_L matrix and the constraint $M_L \succcurlyeq 0$. Thus, $M = (\omega^1)^T \omega^1$. Note that, whenever J contains only one variable x_i (possibly multiple times), this would result in a trivial constraint and, therefore, these constraints are disregarded with one exception: if $|J| = 2$, then ω^1 is replaced with $(1, x_i)$. With this exception, this approach is mathematically equivalent for quadratic problems to the SOCP approach we present in Section 3.3 below.

Approach 2. For each J-set J, ω^2 is used to define the M_L matrix and the constraint $M_L \succcurlyeq 0$.

Although constraints building upon ω^3 lead to tighter relaxations, they generate significantly bigger matrices and increase the complexity of the resulting SDP problems. Because of this, in our setting they are not competitive with Approaches 1 and 2 above, and so we do not include them in the computational analysis of Section 4.

3.3. SOCP CONSTRAINTS

We now describe the second-order cone constraints which, with respect to the SDP ones, lead to looser relaxations but, on the other hand, can be solved more efficiently by state-of-the-art optimization solvers. For each J-set J and each pair of variables present in J, $x_i \neq x_j$, we define the following second-order cone constraint:

$$\frac{X_{ii} + X_{jj}}{2} \geq \left\| \begin{pmatrix} X_{ij} \\ \frac{X_{ii} - X_{jj}}{2} \end{pmatrix} \right\|_2. \tag{2}$$

We argue now why these constraints are valid cuts, *i.e.,* they never remove solutions feasible to (**PO**). Constraint (2) can be equivalently rewritten as $X_{ii}X_{jj} \geq X_{ij}^2$. Then, given a solution of a linear relaxation satisfying the RLT identities in (1), the above condition reduces to $x_i x_i x_j x_j \geq x_i x_j x_i x_j$, which is trivially true. Note that constraints in (2) are trivially true if $i = j$ and, hence, whenever we have a variable x_i appearing twice or more in J, we instead add the second-order constraint

$$\frac{1 + X_{ii}}{2} \geq \left\| \begin{pmatrix} x_i \\ \frac{1 - X_{ii}}{2} \end{pmatrix} \right\|_2,$$

which is equivalent to $X_{ii} \geq x_i x_i$ and, for solutions satisfying (1), is again trivially true.

3.4. BINDING SOCP AND SDP CONSTRAINTS

Since solving SOCP or SDP problems is usually more time-consuming than solving linear programming problems, we define a new approach in order to reduce the time needed for solving the resulting RLT relaxation with SOCP or SDP constraints. This consists of checking which conic constraints (second-order cone or semidefinite) are binding after solving the first relaxation, *i.e.*, this is done only once, at the root node. Thereafter, only these binding constraints are used to tighten future linear relaxations. This approach significantly reduces the number of second-order cone or semidefinite constraints in the relaxations and, although these new relaxations are not as tight, one might expect that the binding constraints at the root node tend to be the most important ones in subsequent relaxations, at least in the first phase of the algorithm. We assess the trade-off between the difficulty of solving the relaxations and how tight they are in the computational analysis in the next section.

4. COMPUTATIONAL RESULTS

4.1. TESTING ENVIRONMENT

All the computational analyses reported in this work have been performed on the supercomputer Finisterrae III, provided by Galicia Supercomputing Centre (CESGA). Specifically, we use nodes powered with 32 cores Intel Xeon Ice Lake 8352Y CPUs with 256GB of RAM connected through an Infiniband HDR network, and 1TB of SSD.

Regarding the datasets, we use three different sets of problems. The first one, DS, is taken from Dalkiran and Sherali (2016) and consists of 180 instances of randomly generated polynomial programming problems of different degrees, number of variables, and density. The second dataset comes from the well known benchmark MINLPLib (Bussieck et al., 2003), a library of Mixed-Integer Nonlinear Programming problems. We have selected from MINLPLib those instances that are polynomial programming problems with box-constrained and continuous variables, resulting in a total of 166 instances. The third dataset comes from another well known benchmark, QPLIB (Furini et al., 2018), a library of quadratic programming instances, for which we made a selection analogous to the one made for MINLPLib, resulting in a total of 63 instances. Hereafter we refer to the first dataset as DS, to the second one as MINLPLib, and to the third one as QPLIB.[1]

We develop our analysis by building upon the global solver for polynomial optimization problems `RAPOSa` (González-Rodríguez et al., 2020). Regarding the auxiliary solvers, `RAPOSa` uses i) `Gurobi` for the linear relaxations ii) `Gurobi` or `Mosek` for the SOCP relaxations, and iii) `Mosek` for the SDP relaxations. The main objective of the thorough numerical analysis developed in this and in the following section is to assess the performance of different SOCP/SDP conic-driven versions of `RAPOSa` with respect to two more traditional ones: basic RLT and RLT with linear SDP-based cuts. More precisely, the full set of ten different versions is as follows:

- `RLT`: standard RLT algorithm (with J-sets).
- `SDP-Cuts`: linear SDP-based cuts added to RLT.
- $\texttt{SOCP}^{G}$: SOCP constraints added to the RLT relaxation and solved with `Gurobi`.
- $\texttt{SOCP}^{G,B}$: same as above, but using only constraints that were binding at the root node.
- $\texttt{SOCP}^{M}$: SOCP constraints added to the RLT relaxation and solved with `Mosek`.
- $\texttt{SOCP}^{M,B}$: same as above, but using only constraints that were binding at the root node.
- $\texttt{SDP}^{1}$: SDP constraints added to the RLT following Approach 1.
- $\texttt{SDP}^{1,B}$: same as above, but using only constraints that were binding at the root node.
- $\texttt{SDP}^{2}$: SDP constraints added to the RLT following Approach 2.
- $\texttt{SDP}^{2,B}$: same as above, but using only constraints that were binding at the root node.

For each instance and each one of the above versions, we run `RAPOSa` with a time limit of one hour.

4.2. NUMERICAL RESULTS AND ANALYSIS

The main goal of the numerical analysis in this section is to show the potential of SOCP/SDP conic-constraints to improve upon the performance of more classic implementations of RLT, such as `RLT` and `SDP-Cuts`. The measures used to evaluate the performance of the different versions of `RAPOSa` are pace^{LB} and npace^{LB}, two performance indicators introduced in Ghaddar et al. (2022) that capture the pace at which a given algorithm closes the gap or, more precisely, the pace at

[1] Instances from DS dataset can be downloaded at https://raposa.usc.es/files/DS-TS.zip, instances from MINLPLib dataset can be downloaded at https://raposa.usc.es/files/MINLPLib-TS.zip, and instances from QPLIB dataset can be downloaded at https://raposa.usc.es/files/QPLIB-TS.zip.

which it increases the lower bound along the branch-and-bound tree. To compute pace^{LB} we use the following formula:

$$\text{pace}^{LB} = \frac{\text{time}}{LB^{\texttt{end}} - LB^{\texttt{root}} + \varepsilon}.$$

Then, npace^{LB} is just a normalized version of pace^{LB} with values in $[0, 1]$. It is computed, for each version of the solver/algorithm, by dividing the best (smallest) pace among all versions to be compared by the pace of the current one.

As thoroughly discussed in Ghaddar et al. (2022), pace^{LB} and npace^{LB} are natural measures that allow to compare the performance of different solvers/algorithms on all the instances of a test set at once, regardless of their difficulty and of how many versions of the underlying solver/algorithm have solved them to optimality. This is different from more common approaches, where the running time is used to evaluate performance on instances solved by all versions, the optimality gap is used for those instances solved by none, and where some decision has to be made regarding those instances solved by some but not all of the versions of the solver/algorithm.

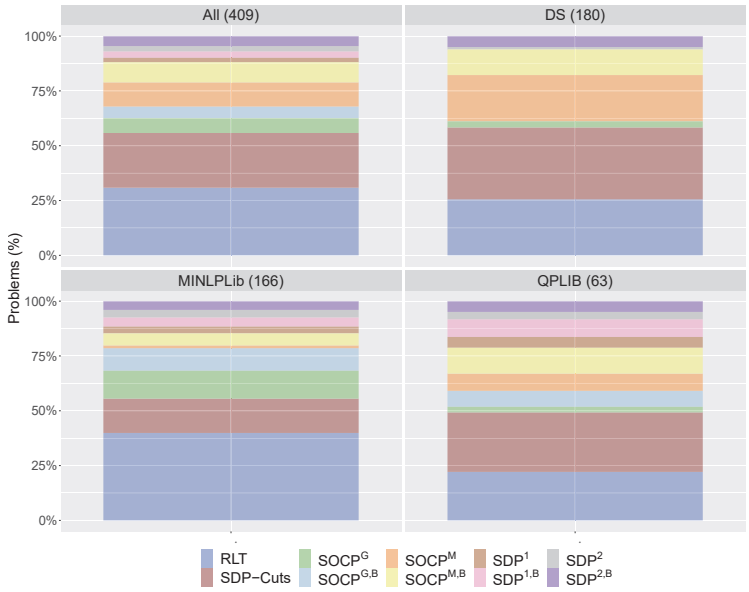


Figure 1: Percentage of instances in which each version delivers the best performance.

Figure 1 shows, for the different sets of problems, the percentage of instances in which each one of the ten versions is the best one. We can see that, quite consistently across the three test sets, a version with either SOCP or SDP constraints is the best one in around 50% of the instances. This is in itself one of the main highlights of this work: RLT versions incorporating nonlinear SOCP/SDP conic constraints can improve the performance of RLT-based algorithms in half of the instances of the sets of problems under consideration.

As seen in Figure 1, the versions with SOCP constraints are the best ones much more often than those with SDP constraints. Regarding the solvers for SOCP versions, `Gurobi` performs notably better in MINLPLib instances, whereas `Mosek` is in most cases better on DS and is also the best one for QPLIB instances. When comparing binding versions with their non-binding counterparts we can see that, for SDP versions, $\texttt{SDP}^{1,B}$ and $\texttt{SDP}^{2,B}$ are the best ones significantly more often than $\texttt{SDP}^{1}$ and $\texttt{SDP}^{2}$, respectively. In the case of SOCP versions, there is no clear winner between binding and non-binding versions. Finally, regarding the RLT versions without conic constraints, `RLT` and `SDP-Cuts`, we can see that each of them turns out to be the best option in around 25% of the instances, with `SDP-Cuts` looking preferable in DS and QPLIB, whereas `RLT` is the best one three times as much in MINLPLib.

Importantly, Figure 1 and the preceding discussion show that there is a lot of variability, with all ten versions showing up as the best choice for a non-negligible percentage of instances. Further, this variability also follows different patterns for the different sets of problems, which motivates

the approach taken in Section 5 below, where we use machine learning techniques to try to learn to choose in advance the most promising RLT version for a given instance.

Figure 2: Percentage of instances in which each version delivers the best performance in high density problems in DS and "water"-related instances in MINLPLib.

A natural question given the results in Figure 1 is whether or not there are specific subclasses of instances in the different test sets where a certain RLT version is noticeably dominant. A deeper analysis of the results shows that this is indeed the case, as presented in Figure 2. First, when looking at instances in DS with high density (larger than 0.5) we can see that the versions relying on SOCP constraints and `Mosek` as a solver, $\texttt{SOCP}^{M}$ and $\texttt{SOCP}^{M,B}$, are the best performing ones in around 70% of the instances; `RLT` goes down to around 10%, `SDP-Cuts` to around 5%, and $\texttt{SDP}^{2}$ and $\texttt{SDP}^{2,B}$ split the remaining 15%. Second, we selected from MINLPLib set all instances whose name contains the word "water", which are problems related to the design of water networks (Castro and Teles, 2013; Teles et al., 2012) and wastewater treatment systems (Castro et al., 2009, 2007). We have that in 15 out of the 28 resulting instances, the best option is one of the following SDP-based ones: $\texttt{SDP}^{1,B}$, $\texttt{SDP}^{2}$, and $\texttt{SDP}^{2,B}$. Again, the instances in which `RLT` or `SDP-Cuts` are the best option are less than 20%. The reasons behind the good performance of the versions based on second order cone constraints for high density problems in DS and of those based on positive semidefinite constraints for "water"-related instances in MINLPLib are definitely worth studying more deeply, but such an analysis is beyond the scope of this work.

	All	**DS high density**	**water**
RLT (across all instances)	12.69	0.82	33994.03
Optimal version (instance by instance)	6.28	0.21	2546.88
Improvement	50.5%	74.4%	92.5%

Table 1: Average values for pace^{LB}.

Despite the results shown in Figure 1 and Figure 2, it is important to ensure that the variability is not spurious. For instance, it might be that all RLT versions performed very similarly to one another, which would turn most of the above discussion meaningless. In Table 1 we present a concise summary of the geometric mean of pace^{LB} for the complete set of instances and for the two special subclasses identified above. The first row measures the performance of `RLT` while the second row measures the performance of a hypothetical RLT version capable of choosing the best performing RLT version in each and every instance. The first column shows that, on aggregate on the whole set of instances, this hypothetical and optimal version would divide the pace by two, *i.e.*, a substantial improvement of 50.5%. This improvement is much more pronounced for high density problems in DS, in which the pace gets divided by four (74.4% improvement), and even more so for "water"-related instances in MINLPLib where the pace becomes more than ten times smaller (92.5% improvement).

5. MACHINE LEARNING FOR DIFFERENT CONIC CONSTRAINTS

In this section we want to exploit the wide variability in the performance of the different RLT versions shown above to try to learn in advance which one should be chosen for a given instance. The goal is to design a machine learning procedure that can be trained using the different features of the instances and then choose the most promising RLT version when confronted with a new instance. The performance of the hypothetical "Optimal version" in Table 1 represents an upper bound on the improvement that can be attained by such a machine learning version.

We follow the framework in Ghaddar et al. (2022), where the authors use learning techniques to improve the performance of the RLT-based solver `RAPOSa` by learning to choose between different branching rules. The improvements reported there are substantial, with the machine learning version delivering improvements of up to 25% with respect to the best original branching rule. They use different features to capture diverse characteristics of each instance and are a key ingredient of the machine learning framework, which consists of predicting the performance (npace^{LB}) of each branching rule on a new instance based on a regression analysis of its performance on the training instances. Then, the rule with a highest predicted performance for the given instance is chosen.

Table 2 shows the remarkable improvement obtained with the machine learning (ML) version of RLT that chooses, for each given instance, the most promising version of the 10-version portfolio. Considering all instances, the ML version improves 32.9% with respect to `RLT`, being the upper bound for learning 50.5%. In instances with high densities from DS test set, it improves a remarkable 69.5% out of the optimal 74.4%, and in "water"-related instances in MINLPLib it improves 56.7% out of 92.5%. It is worth noting that the size of these improvements is comparable, and even superior, to those obtained in Ghaddar et al. (2022) when this learning scheme was introduced to learn to choose between branching rules, which shows the robustness of the proposed approach.

	All	**DS high density**	**water**
RLT (across all instances)	12.69	0.82	33994.03
ML-based version	8.51	0.25	14727.22
Optimal version (instance by instance)	6.28	0.21	2546.88
Improvement after learning	32.9%	69.5%	56.7%
Optimal improvement (upper bound for learning)	50.5%	74.4%	92.5%

Table 2: Performance of the ML-based version with respect to pace^{LB} (average across test sets).

Figure 3 represents, side by side, how often each of the ten RLT versions is selected by the ML version and by the optimal one. We can see that the behavior of the former mimics quite well that of the latter in the three sets of instances. Given that the learning is conducted jointly on the whole set of instances, the fact that the ML version adapts to the instances in DS, MINLPLib, and QPLib, is reassuring about the quality of the learning process. In particular, the SOCP versions with `Mosek` are primarily chosen for the DS test set and the SDP versions are mainly chosen for the QPLIB test set, the one in which they are optimal more often. In Figure 4 we further explore the behavior for the high density problems in DS and for "water"-related instances in MINLPLib. We see that, again, the ML version mimics the patterns of the optimal version. The dominant version in DS, $\texttt{SOCP}^M$, is chosen in almost 75% of the instances. Similarly, the three SDP dominant versions in "water"-related instances are chosen almost 50% of the time. The dominant version in DS, $\texttt{SOCP}^M$, is chosen in almost 75% of the instances.

Figure 5 presents boxplots summarizing the performance according to npace^{LB} for all instances and for each test set. Recall that, by definition, values close to one mean that the corresponding version is almost the best one, whereas values close to zero imply that its pace is much worse than the best one. We can see that, although `RLT` and `SDP cuts` versions are, on aggregate, the best ones in all three test sets, they are significantly outperformed by the ML version in the three of them. The improvement is particularly noticeable in DS, where the learning criterion is, by far, better than all underlying versions.

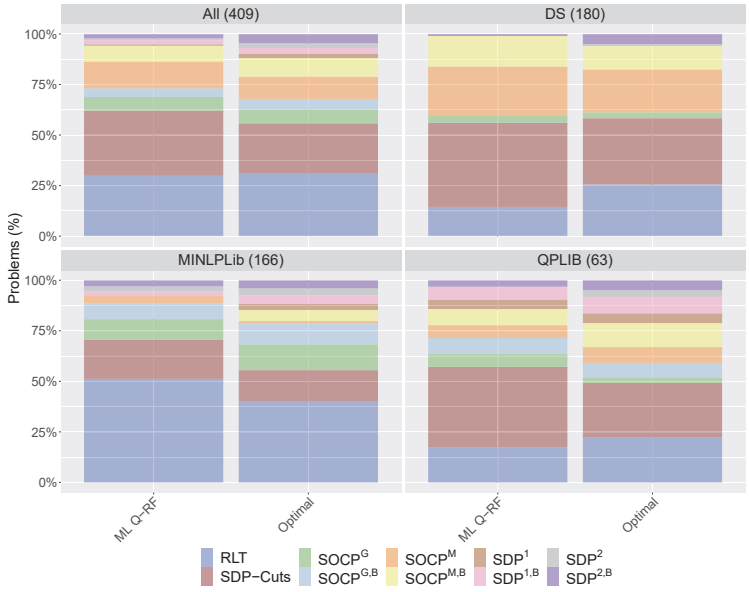


Figure 3: Percentages in which each version is selected by the ML version and by the optimal one.

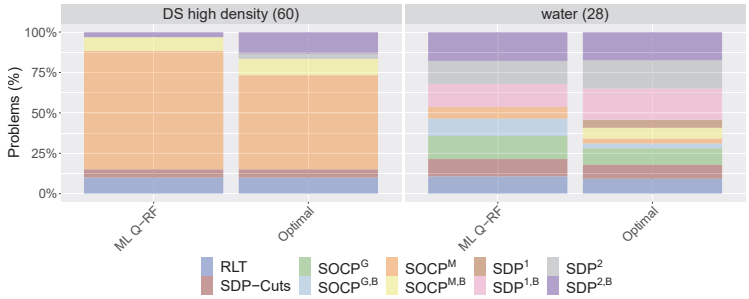


Figure 4: Percentages in which each version is selected by the ML version and by the optimal one in high density problems in DS and "water"-related instances in MINLPLib.

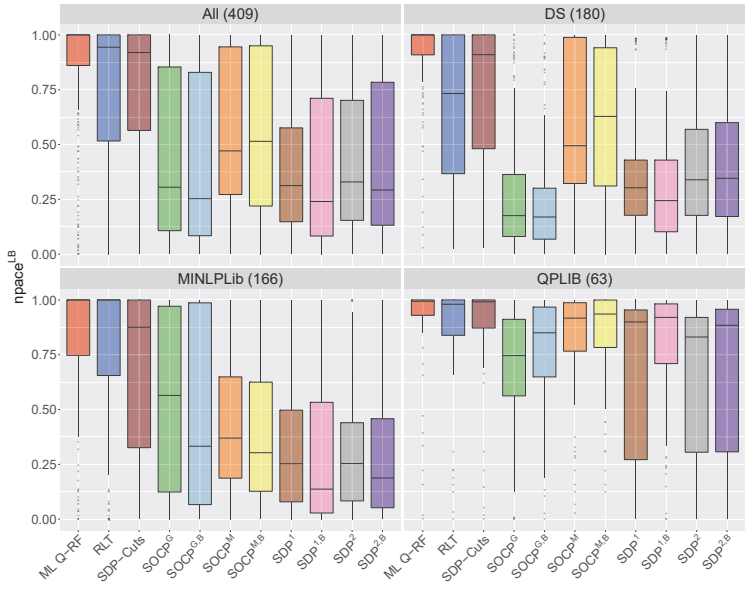


Figure 5: Boxplot of npace^{LB} for each approach.

6. CONCLUSIONS AND FUTURE RESEARCH

The main contribution of this work is to show that the solution to global optimality via branch-and-bound schemes of non-convex optimization problems and, in particular, polynomial optimization ones, can benefit from tightening the underlying relaxations with conic constraints. We explore different families of such constraints, building upon either second-order cones or positive semidefiniteness. We also show that the potential of these conic constraints can be successfully exploited by embedding them into a learning framework. The main goal is to predict which is the most promising type of constraint to add to the RLT relaxation at each node of the underlying branch-and-bound algorithm when confronted with a new instance. The results in Section 4 show that the versions with SOCP/SDP conic constraints deliver consistently good results for instances in specific subclasses of problems: high density problems in DS and of those based on positive semidefinite constraints for "water"-related instances in MINLPLib.

As a future step, one may wonder to what extent one might get an even superior performance if the learning analysis was further specialized for the current setting: for example including features capturing some "conic" characteristics of the polynomial optimization problems and fine-tuning the regression techniques. Furthermore, an important direction for future research is to improve the understanding on the structure of these problems and the specificities that lead to better performance of SOCP and SDP constraints, respectively. Another direction is to investigate the number of SOCP and SDP constraints added at different nodes of the branch-and-bound tree. Additionally, we aim to extend this framework for various relaxations of polynomial optimization problems besides RLT.

REFERENCES

Andersen, E. D. and Andersen, K. D. (2000). *The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm*. Springer US, Boston, MA.

Baltean-Lugojan, R., Bonami, P., Misener, R., and Tramontani, A. (2019). Scoring positive semidefinite cutting planes for quadratic optimization via trained neural networks. Technical report, Optimization-online 7942.

Bengio, Y., Lodi, A., and Prouvost, A. (2021). Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421.

Bonami, P., Lodi, A., Schweiger, J., and Tramontani, A. (2019). Solving quadratic programming by cutting planes. *SIAM Journal on Optimization*, 29(2):1076–1105.

Buchheim, C. and Wiegele, A. (2013). Semidefinite relaxations for non-convex quadratic mixed-integer programming. *Mathematical Programming*, 141(1):435–452.

Burer, S. and Vandenbussche, D. (2008). A finite branch-and-bound algorithm for nonconvex quadratic programming via semidefinite relaxations. *Mathematical Programming*, 113(2):259–282.

Burer, S. and Ye, Y. (2020). Exact semidefinite formulations for a class of (random and non-random) nonconvex quadratic programs. *Mathematical Programming*, 181(1):1–17.

Bussieck, M. R., Drud, A. S., and Meeraus, A. (2003). MINLPLib-a collection of test models for mixed-integer nonlinear programming. *INFORMS Journal on Computing*, 15:114–119.

Castro, P. M., Matos, H. A., and Novais, A. Q. (2007). An efficient heuristic procedure for the optimal design of wastewater treatment systems. *Resources, conservation and recycling*, 50(2):158–185.

Castro, P. M. and Teles, J. P. (2013). Comparison of global optimization algorithms for the design of water-using networks. *Computers & chemical engineering*, 52:249–261.

Castro, P. M., Teles, J. P., and Novais, A. Q. (2009). Linear program-based algorithm for the optimal design of wastewater treatment systems. *Clean Technologies and Environmental Policy*, 11(1):83–93.

Dalkiran, E. and Sherali, H. D. (2013). Theoretical filtering of RLT bound-factor constraints for solving polynomial programming problems to global optimality. *Journal of Global Optimization*, 57(4):1147–1172.

Dalkiran, E. and Sherali, H. D. (2016). RLT-POS: Reformulation-linearization technique-based optimization software for solving polynomial programming problems. *Mathematical Programming Computation*, 8:337–375.

Elloumi, S. and Lambert, A. (2019). Global solution of non-convex quadratically constrained quadratic programs. *Optimization methods and software*, 34(1):98–114.

FICO (2022). FICO Xpress Optimization Suite. Available at: https://www.fico.com/en/products/fico-xpress-optimization.

Furini, F., Traversi, E., Belotti, P., Frangioni, A., Gleixner, A., Gould, N., Liberti, L., Lodi, A., Misener, R., Mittelmann, H., Sahinidis, N., Vigerske, S., and Wiegele, A. (2018). QPLIB: a library of quadratic programming instances. *Mathematical Programming Computation*, 1:237–265.

Ghaddar, B., Gómez-Casares, I., González-Díaz, J., González-Rodríguez, B., Pateiro-López, B., and Rodríguez-Ballesteros, S. (2022). Learning for spatial branching: An algorithm selection approach. Technical report, .

Ghaddar, B., Vera, J. C., and Anjos, M. F. (2011). Second-order cone relaxations for binary quadratic polynomial programs. *SIAM Journal on Optimization*, 21(1):391–414.

González-Rodríguez, B., Ossorio-Castillo, J., González-Díaz, J., González-Rueda, Á. M., Penas, D. R., and Rodríguez-Martínez, D. (2020). Computational advances in polynomial optimization: RAPOSa, a freely available global solver. Technical report, Optimization-online 7942.

Gurobi Optimization (2022). Gurobi Optimizer Reference Manual. Available at: http://www.gurobi.com.

IBM Corp. (2022). IBM ILOG CPLEX Optimization Studio. CPLEX User's Manual. Available at: https://www.ibm.com/es-es/products/ilog-cplex-optimization-studio.

Kannan, R., Nagarajan, H., and Deka, D. (2022). Learning to accelerate partitioning algorithms for the global optimization of nonconvex quadratically-constrained quadratic programs. *arXiv preprint arXiv:2301.00306*.

Lodi, A. and Zarpellon, G. (2017). On learning and branching: a survey. *Top*, 25(2):207–236.

MOSEK ApS (2022). *Introducing the MOSEK Optimization Suite 9.3.20*.

Sherali, H. D., Dalkiran, E., and Desai, J. (2012). Enhancing RLT-based relaxations for polynomial programming problems via a new class of v-semidefinite cuts. *Computational Optimization and Applications*, 52(2):483–506.

Sherali, H. D. and Tuncbilek, C. H. (1992). A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *Journal of Global Optimization*, 2(1):101–112.

Shor, N. Z. (1987). An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics*, 23(5):695–700.

Teles, J. P., Castro, P. M., and Matos, H. A. (2012). Global optimization of water networks design using multiparametric disaggregation. *Computers & Chemical Engineering*, 40:132–147.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

SELECCIÓN GENÓMICA EN ALTA DIMENSIÓN: CUANDO HAY MÁS COVARIABLES QUE MUESTRAS

Laura Freijeiro-González[1], Manuel Febrero-Bande[1] y Wenceslao González-Manteiga[1]

[1]Centro de Investigación y Tecnología Matemática de Galicia (CITMAga). Departamento de Estadística, Análisis Matemático y Optimización. Universidade de Santiago de Compostela.

RESUMEN

En la selección genómica es muy común encontrarse en un contexto de alta dimensión donde se tienen más covariables o genes que muestras ($p > n$). Es en esta situación donde los procesos usuales de selección de covariables comienzan a funcionar mal. En este trabajo se motivará la necesidad de recurrir a procedimientos adecuados de selección adaptados al caso $p > n$ para poder tratar con estas situaciones. Para ello, se motivará la selección genética en datos de producción de riboflavina de la bacteria *Bacillus subtilis*. Se revisarán técnicas diseñadas para este contexto específico y se motivará la importancia de escoger adecuadamente un procedimiento de selección en relación con la estructura de dependencia y la escala de las covariables. Para ello, se propondrán ciertas pautas de decisión resultantes de estudios realizados. Se ilustrará cómo proceder en la práctica usando el conjunto de datos de riboflavina y se extraerán algunas conclusiones generales.

Palabras y frases clave: Alta dimensión; correlación de distancias; métodos de penalización; modelo de regresión lineal; selección de covariables

1. INTRODUCCIÓN Y MOTIVACIÓN

En esta era de la información, cada vez disponemos de una mayor cantidad de datos a causa de los avances informáticos y tecnológicos. La abundancia de datos puede incluso sobrepasar la capacidad de la metodología estadística clásica. Como resultado, tiene especial interés el desarrollo y dominio de nuevos procedimientos estadísticos adaptados a las demandas actuales. De esta forma, se persiguen metodologías que sean capaces de tener en cuenta toda la información disponible que sea útil, desechando aquella que solo aporte ruido.

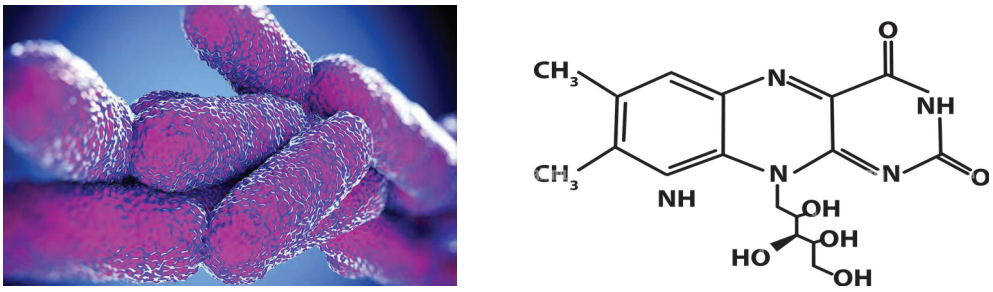


Figura 1: Izquierda: Bacterias *Bacillus subtilis*. Derecha: Formulación riboflavina/vitamina B2.

Ejemplos de este fenómeno pueden encontrarse en la rama sanitaria o en la biología, donde se estudian minuciosamente las características de un individuo u organismo para el desarrollo de tratamientos personalizados o para conseguir explicar ciertos comportamientos. Con especial interés se destacan los estudios genómicos, donde lo que se busca es detectar qué genes afectan a ciertas funciones fisiológicas o cómo estos contribuyen a aumentar la predisposición a sufrir una

determinada enfermedad. Esto contribuye a un mejor entendimiento de los procesos vitales y ayuda a la prevención de enfermedades, así como a la creación de tratamientos personalizados para cada paciente. En estos estudios es muy común encontrar que el número de covariables consideradas, p, excede al tamaño muestral disponible, n. En este contexto de alta dimensión, considerando $p > n$, los procedimientos usuales comienzas a funcionar mal o no son aplicables.

Un ejemplo es el estudio de los genes que afectan a la producción de riboflavina o vitamina B2 de la bacteria *Bacillus subtilis* (ver Figura 1). Entre otras funciones, esta vitamina es la encargada de la respiración celular del cuerpo y tiene propiedades "antienvejecimiento". Se encuentra en alimentos como los huevos, los vegetales verdes o la leche. Con el fin de explicar la producción de riboflavina o vitamina B2 de esta bacteria, Bühlmann *et al.* (2014) miden la expresión génica de un total de $p = 4088$ genes en $n = 71$ experimentos. Como resultado, se busca detectar qué genes de los estudiados son realmente importantes en la producción de riboflavina y cómo influyen en la misma en un contexto de alta dimensión donde $p > n$.

Asumiendo que existe una relación lineal entre una variable de interés o respuesta, Y, y p covariables explicativas $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$, todas ellas centradas, dicha relación puede modelarse a través del modelo de regresión lineal que viene dado por

$$Y = X\beta + \varepsilon, \tag{1}$$

donde ε es el error del modelo, con $\mathbb{E}[\varepsilon] = 0$ y $\mathbb{E}[\varepsilon^2] = \sigma^2$, y $\beta \in \mathbb{R}^p$ es un vector a estimar.

Considerando $(\mathbf{X_n}, \mathbf{Y_n}) = \{(x_i, y_i), i = 1, \ldots, n\}$ una muestra aleatoria e idénticamente distribuida de la función de distribución conjunta de $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, se puede obtener una estimación de β por el método de mínimos cuadrados. Este se basa en la resolución del problema

$$\min_\beta \phi(\beta) = \min_\beta \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = \min_\beta \ (\mathbf{Y_n} - \mathbf{X_n}\beta)^\top (\mathbf{Y_n} - \mathbf{X_n}\beta) = \min_\beta \ \|\mathbf{Y_n} - \mathbf{X_n}\beta\|_n^2, \tag{2}$$

siendo $\|\cdot\|_n$ la norma euclidiana en $\mathbb{R}^n$.

Derivando e igualando a cero la expresión $\phi(\beta)$ dada en (2) se llega a que el estimador de mínimos cuadrados viene dado por

$$\hat{\beta} = \left(\mathbf{X_n}^\top \mathbf{X_n}\right)^{-1} \mathbf{X_n}^\top \mathbf{Y_n}. \tag{3}$$

Sin embargo, en una situación donde $p > n$, no es posible obtener el estimador $\hat{\beta}$ de (3). Esto se debe a que $\mathbf{X_n}$ es una matriz de dimensión $n \times p$, $\mathbf{X_n}^\top \mathbf{X_n}$ es una matriz $p \times p$ dimensional, y el Corolario 1 garantiza que $\text{rank}(\mathbf{X_n}^\top \mathbf{X_n}) \leq n < p$, siendo $\text{rank}(\cdot)$ el operador rango de una matriz. Como sabemos que

$$\exists (\mathbf{X_n}^\top \mathbf{X_n})^{-1} \ \Leftrightarrow \ \det(\mathbf{X_n}^\top \mathbf{X_n}) \neq 0 \ \Leftrightarrow \ \text{rank}(\mathbf{X_n}^\top \mathbf{X_n}) = p,$$

donde $\det(\cdot)$ denota el operador determinante, no se puede garantizar la existencia de la inversa de $(\mathbf{X_n}^\top \mathbf{X_n})^{-1}$ de forma única. Por lo tanto, no habría unicidad del estimador de mínimos cuadrados.

Corolario 1 *Siendo A una matriz de dimensión $p \times n$ y B una matriz de dimensión $n \times p$ donde $p > n$, entonces*

$$\left.\begin{array}{c} \text{rank}(A \cdot B) \leq \text{rank}(A) \\ y \\ \text{rank}(A \cdot B) \leq \text{rank}(B) \end{array}\right\} \Rightarrow \text{rank}(A \cdot B) \leq n$$

dado que $\text{rank}(A) \leq n$ *y* $\text{rank}(B) \leq n$ *porque* $p > n$.

Además, cuando $p > n$, aparecen otros problemas como la maldición de la dimensionalidad o posibles efectos de colinearidad. Ver, por ejemplo, Hastie *et al.* (2009), Giraud (2014) o Hastie *et al.* (2015). Todos estos inconvenientes también se heredan para formulaciones más complejas del modelo de regresión, sin asumir linealidad como en (1).

Como resultado, en un contexto de regresión en alta dimensión con $p > n$, como es el caso de muchos estudios genómicos, tiene especial interés aplicar un paso preliminar de reducción de

la dimensión. Esto permite incluir únicamente aquellas covariables que contengan información relevante para el modelo, excluyendo aquellas que son innecesarias o redundantes y, por lo tanto, considerar menos de p términos. De esta forma se le hace frente a los problemas citados previamente. Para este fin, se puede recurrir a métodos de penalización, como la famosa regresión LASSO (Tibshirani (1996)), o aplicar técnicas de selección de covariables empleando coeficientes como la correlación de distancias de Székely *et al.* (2007).

El documento se organiza como sigue. En la Sección 2 se introducen métodos de selección de covariables para el contexto $p > n$, diferenciando dos casos: asumiendo o no cierta estructura del modelo de regresión. Posteriormente, en la Sección 3, se estudia la importancia de considerar las estructuras de dependencia y las diferencias de escalas que pueden existir entre las covariables en el proceso de selección. En la Sección 4 se muestra un ejemplo de aplicación a un contexto genómico, empleando los datos de riboflavina introducidos previamente. Finalmente, se arrojan algunas conclusiones generales en la Sección 5.

2. MÉTODOS DE SELECCIÓN DE COVARIABLES

A la hora de seleccionar covariables en un modelo de regresión, como el dado en la formulación (1), uno podría pensar en implementar un criterio de tipo AIC (Akaike (1998)) o BIC (Schwarz (1978)). Estos añaden una restricción al problema de mínimo cuadrados (ver ejemplo en (2)), restringiendo el número de covariables incluidas en el modelo. Sin embargo, estos enfoques son muy costosos computacionalmente, ya que habría que realizar 2^p comparaciones para p covariables. Por lo tanto, este enfoque no es adecuado cuando p es grande. Además, en el contexto de alta dimensión con $p > n$, estos criterios tienden a sobreajustar los resultados y a seleccionar términos que no son relevantes. Ver Giraud *et al.* (2012) para más detalle. En conclusión, los criterios basados en información generalizada no son adecuados para nuestro contexto de interés. Una filosofía similar, también basada en penalizar el número de variables que se incluyen en el modelo, son los métodos de penalización. Estos se introducen a continuación en la Sección 2.1.

Por otro lado, se podría pensar en hacer una selección de covariables previa al ajuste del modelo. De esta forma, se podría recurrir a coeficientes de dependencia para ver cuáles de las p covariables $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ son las que están más relacionadas con la respuesta Y, incluyendo únicamente aquellas que sean significativamente relevantes. Para este fin se podría emplear el famoso coeficiente de correlación (Pearson (1920)) o alternativas más robustas como el coeficiente de Spearman (Wissler (1905)) o la τ de Kendall (Kendall (1938)). Sin embargo, el coeficiente de correlación solo es capaz de detectar relaciones lineales y el coeficiente de Spearman, junto con la τ de Kendall, solo funcionan para patrones monótonos. Una opción más global es la propuesta de Székely *et al.* (2007) con la correlación de distancias. Esta permite detectar cualquier tipo de estructura de dependencia. Este coeficiente se estudia en la Sección 2.2.

2.1 Asumiendo una cierta estructura: Métodos de penalización

Los métodos de penalización necesitan de la asunción de una cierta estructura en el modelo, como la linealidad. Estos añaden una función penalizadora en el proceso de estimación, penalizando el número de covariables incluidas, y permiten seleccionar covariables a la vez que se estima el modelo. En el caso del modelo lineal introducido en (1), se añade una función $p_\lambda(\beta)$ al problema de mínimos cuadrados visto en (2). Esta depende de una parámetro regularizador $\lambda > 0$ y penaliza el vector de coeficientes β. Una forma de tomar la función penalizadora es considerar $p_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$, la cual es de tipo L_1 y garantiza la convexidad del problema. Esta implementación da lugar a la famosa y ampliamente estudiada regresión LASSO de Tibshirani (1996). Se puede ver una retrospectiva de la misma en Tibshirani (2011). Por lo tanto, ahora se necesitaría resolver el problema

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{4}$$

resultando en un vector $\hat{\beta}$ *sparse* con muchas componentes nulas. Esto permite llevar a cabo selección de covariables, entrando en juego únicamente aquellas cuyo β_j asociado sea no nulo; así como estimar el vector β del modelo simultáneamente, incluso en el caso de $p > n$. Además, la convexidad del problema se traduce en ventajas a la hora de encontrar una solución óptima de (4).

A pesar de las buenas propiedades citadas, la regresión LASSO cuenta con importantes inconvenientes en la práctica. Estos inconvenientes están relacionados con la naturaleza sesgada del estimador $\hat{\beta}^{LASSO}$ (ver Capítulo 3 de Hastie *et al.* (2009), Capítulo 4 de Giraud (2014) o Capítulo 2 de Hastie *et al.* (2015)), así como con la necesidad de garantizar que se cumplen ciertas condiciones teóricas que no son fáciles de verificar en la práctica (Bunea (2008), Lounici (2008), Bühlmann y Van De Geer (2011)), como la *irrepresentable condition* (Zhao y Yu (2006), Zou (2006), Yuan y Lin (2007)). Existe además una relación inversa entre la proporción de descubrimientos falsos (FDP) y la proporción de verdaderos positivos (TPP). Esto se traduce en el hecho de que la regresión LASSO tiende a incluir muchas covariables irrelevantes en su selección para garantizar que incorpora todas las importantes. Ver Wasserman y Roeder (2009) o Su *et al.* (2017) para más información. Otra inquietud es como seleccionar adecuadamente el parámetro de regularización λ. En la práctica, un valor óptimo de λ necesita conocer de antemano la varianza del error del modelo σ^2 (Bühlmann y Van De Geer (2011)). Como esto último no es posible, es común recurrir a técnicas de validación cruzada o usar criterios de información generalizada como el BIC (Homrighausen y McDonald (2018)). Como se comentó al comienzo de la Sección 2, esta última opción únicamente funcionará bien en el contexto clásico de $n \geq p$. Con el fin de solucionar algunos de estos problemas, se han propuesto nuevas modificaciones y alternativas en la literatura. Algunos de estos procedimientos se muestran en la Tabla 2 al final del documento.

2.2 Sin asumir estructura: Correlación de distancias

En el caso de que no se quiera realizar ninguna suposición sobre la estructura del modelo, se puede recurrir a coeficientes de correlación para implementar un primer paso de selección de covariables. De esta forma, se testaría como de relevante es cada una de las $X_1, \ldots, X_p$ covariables en relación con la respuesta Y. Posteriormente, se incluirían en el modelo únicamente aquellos términos que han demostrado que su correlación asociada es significativamente grande en base a algún criterio.

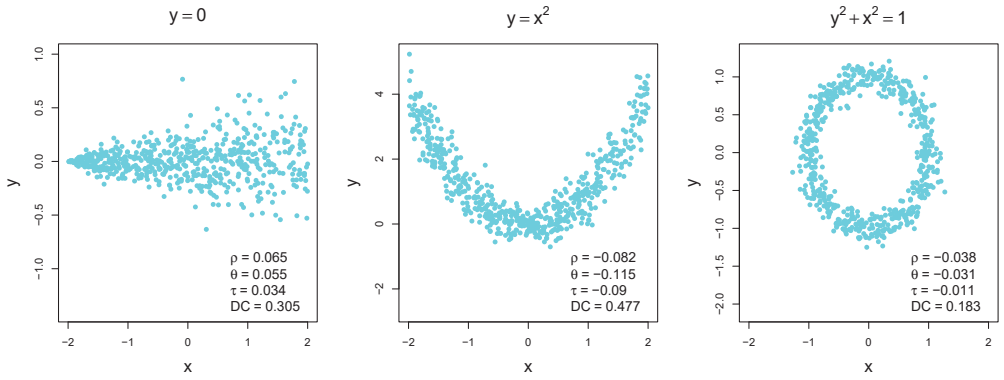


Figura 2: Valor del coeficiente de correlación (ρ), del coeficiente de Spearman (θ), de la τ de Kendall (τ) y de la correlación de distancias (DC) para diferentes escenarios.

Dentro de las medidas clásicas de dependencia encontramos el coeficiente de correlación (Pearson (1920)) que es capaz de detectar solo relaciones lineales. Otras alternativas son el coeficiente de Spearman (Wissler (1905)) o la τ de Kendall (Kendall (1938)). Estas propuestas son robustas a la presencia de atípicos y detectan cualquier tipo de relación monótona. Sin embargo, como se comentaba al inicio del trabajo en la Sección 1, cada vez nos encontramos con más datos, resultando en relaciones muy diversas. Por lo tanto, es interesante ser capaz de detectar también relaciones que no sean monótonas. Székely *et al.* (2007) introduce el coeficiente de correlación de distancias (DC) como solución a este problema. Una comparativa entre estos coeficientes puede verse en la Figura 2 para un escenario heterocedástico y dos no monótonos. En vista de los resultados de la Figura 2, se puede apreciar que la DC obtiene siempre los valores más próximos a la unidad. Esto ilustra su buen funcionamiento cuando no se cumplen los requisitos de linealidad o monotonía.

La DC es una nueva medida de dependencia que, dados dos vectores aleatorios $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ con $p, q \geq 1$, detecta cualquier tipo de relación entre ambos. Para ello, testa la hipótesis H_0: $X \perp Y$, donde $X \perp Y$ denota la independencia entre ambos vectores. Esta condición de

independencia se puede reescribir a través del contraste

$$H_0 : \varphi_{X,Y} = \varphi_X \varphi_Y \quad \text{vs.} \quad H_1 : \varphi_{X,Y} \neq \varphi_X \varphi_Y \tag{5}$$

donde $\varphi_{X,Y}$ es la función característica conjunta y φ_X, φ_Y son las versiones marginales de X, Y.

Para determinar si hay pruebas para rechazar la hipótesis nula de independencia de (5), Székely *et al.* (2007) proponen como estadístico la covarianza de distancias dada por

$$DCov^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{\|t\|_p^{p+1} \|s\|_q^{q+1}} dt\, ds, \tag{6}$$

donde $c_r = \frac{\pi^{(r+1)/2}}{\Gamma((r+1)/2)}$ para $r \geq 1$, siendo $\Gamma(\cdot)$ la función gamma, $|f| = f\bar{f}$ es la norma compleja para una función $f \in \mathbb{C}$ y $\|\cdot\|_r$ es la norma euclidiana en $\mathbb{R}^r$.

El coeficiente $DCov$ introducido en (6) está bien definido siempre que se garantice que $\mathbb{E}[\|X\|_p] < \infty$ y $\mathbb{E}[\|Y\|_q] < \infty$. Ver Székely *et al.* (2007) o Szekely y Rizzo (2017) para más detalles. Además, este verifica siempre que $DCov \geq 0$ y $DCov = 0$ si y solo si $X \perp Y$.

Normalizando el coeficiente de covarianza de distancias introducido en (6) se puede obtener la correlación de distancias dada por

$$DC^2(X,Y) = \begin{cases} \frac{DCov^2(X,Y)}{\sqrt{DCov^2(X,X)DCov^2(Y,Y)}}, & DCov^2(X,X)DCov^2(Y,Y) > 0, \\ 0, & DCov^2(X,X)DCov^2(Y,Y) = 0. \end{cases} \tag{7}$$

Este término verifica que $0 \leq DC(X,Y) \leq 1$, y $DC(X,Y) = 0$ si y solo si $X \perp Y$.

En Székely *et al.* (2007) y Szekely y Rizzo (2017) se pueden ver varias propiedades de estos coeficientes así como la obtención de su distribución asintótica. Además, proponen estimadores empíricos de ambas cantidades para testar (5) en la práctica. Dada $(\mathbf{X_n}, \mathbf{Y_n}) = \{(X_i, Y_i), i = 1, \ldots, n\}$ una muestra aleatoria e idénticamente distribuida de $(X,Y) \in \mathbb{R}^p \times \mathbb{R}^q$ definen $A_{il} = a_{il} - \bar{a}_{i\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ por medio de las cantidades

$$a_{il} = \|X_i - X_l\|_p, \quad \bar{a}_{i\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{il}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{i=1}^{n} a_{il} \quad \text{y} \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,l=1}^{n} a_{il}, \tag{8}$$

y de forma similar para $B_{il} = b_{il} - \bar{b}_{i\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ con $b_{il} = \|Y_i - Y_l\|_q$. Entonces, la covarianza de distancia empírica al cuadrado $DCov_n^2(\mathbf{X_n}, \mathbf{Y_n})$, siendo el estimador empírico de (6), es la cantidad no negativa definida por

$$DCov_n^2(\mathbf{X_n}, \mathbf{Y_n}) = \frac{1}{n^2}\sum_{i,l=1}^{n} A_{il}B_{il}. \tag{9}$$

De forma similar se construye la versión empírica para la DC enchufando estas versiones empíricas en la formula (7).

En conclusión, siendo capaces de medir las distancias entre las muestras de los vectores X e Y, recogidas respectivamente en las matrices A y B, se pueden obtener las versiones empíricas de $DCov$ y DC y, por lo tanto, implementar el contraste de independencia propuesto en (5). El uso de la DC permite llevar a cabo selección de covariables mediante procedimientos iterativos. Un ejemplo es el método DC.VS propuesto por Febrero-Bande *et al.* (2019).

3. LA IMPORTANCIA DE LA DEPENDENCIA Y LA ESCALA

En esta sección ilustraremos la importancia de tener en cuenta las estructuras de dependencia y el fenómeno de la escala presentes en las covariables al aplicar las técnicas vistas en la Sección 2. Esto se hará a través de estudios de simulación en el marco del modelo de regresión lineal introducido en (1). Determinando, de esta forma, las mejores opciones en cada caso.

3.1 Covariables con estructura de dependencia

A la hora de seleccionar covariables es necesario estudiar si existe algún tipo de estructura de dependencia entre ellas: si las covariables están todas relacionadas entre sí, si solo lo están algunas

de ellas, etc. Un estudio detallado acerca del efecto de la dependencia en la selección de covariables cuando $p > n$ puede verse en Freijeiro-González *et al.* (2022).

Uno de los escenarios que se consideran en dicho estudio es el **Escenario 3.a: Covarianza unitaria tipo Toeplitz**, el cual asume que todas las covariables están relacionadas entre sí a través de una estructura de dependencia tipo Toeplitz. De esta forma, se simula el modelo (1) donde $X \in N(0, \Sigma)$, siendo $(\Sigma)_{jk} = \sigma_{jk} := \rho^{|j-k|}$ con $\rho = 0{,}5, 0{,}9$ para $j, k = 1, \ldots, p$, y $\varepsilon \in N_n(0, \sigma^2 I_n)$. Se toma $p = 100$, se consideran un total de $s = 15$ ($s < p$) covariables relevantes, localizadas en las primeras 15 posiciones, y se testa con distintos tamaños muestrales para n.

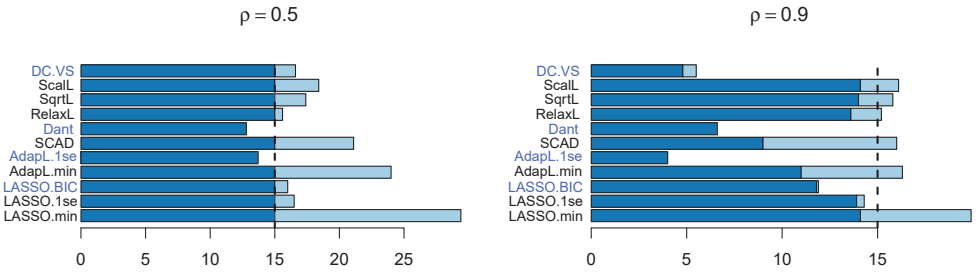


Figura 3: Comparativa del número de variables relevantes seleccionadas (área oscura) frente a las ruidosas (área clara) para $n = 400$ en el Escenario 3.a. La línea discontinua marca el valor $s = 15$.

Los resultados de la selección de covariables hecha por los diferentes procedimientos introducidos en las Sección 2 pueden verse en las Figura 3. Se aprecia que no todos los procedimientos tienden a recuperar las $s = 15$ covariables, especialmente cuando la dependencia es fuerte ($\rho = 0{,}9$), y que además, muchos tienen a añadir ruido en el proceso. En este caso, podemos diferenciar entre algoritmos que hacen uso de la estructura de dependencia y detectan que no son necesarias las $s = 15$ covariables importantes (en azul), reduciendo la tasa de falsos positivos ($FDP \downarrow$), y otros que buscan recuperar estas en su totalidad, aumentando la tasa de verdaderos positivos ($TPP \uparrow$). Esto también se extiende a otros escenarios de dependencia. Por lo tanto, la elección de un procedimiento adecuado dependerá del tipo de dependencia y de nuestro objetivo. Más detalles se pueden consultar en Freijeiro-González *et al.* (2022).

3.2 Covariables en distintas escalas

Similar al estudio de dependencia, también es necesario tener en cuenta si las covariables se encuentran en distintas escalas. En este caso, será necesario aplicar un primer paso de estandarización. Veremos qué ocurre con los distintos procedimientos considerados en las Sección 2 al no aplicar estandarización y cuando se estandarizan las covariables. Un estudio detallado acerca del efecto de la escala se puede ver en el Capítulo 3.2 de Freijeiro-González (2023).

A modo ilustrativo, consideramos el **Escenario 1: Independencia con distintas escalas** del Capítulo 3.2.1 de Freijeiro-González (2023). Este es similar al escenario introducido anteriormente en la Sección 3.1, pero ahora se consideran $s = 10$ covariables importantes localizadas en las primeras posiciones y la matriz Σ es diagonal. En particular, veremos lo que ocurre en el **Escenario 1.b** tomando $\text{diag}(\Sigma^{1.b}) = (0{,}5, 0{,}5, 1, 1, 3, 3, 10, 10, 25, 25, 1, \overset{p-s)}{\ldots}, 1)$ y en el **Escenario 1.c** con $\text{diag}(\Sigma^{1.c}) = \left((\text{diag}(\Sigma^{1.b})_j)_{j=1}^{s}, 0{,}5, 0{,}5, 0{,}5, 1, 1{,}5, 1, 1{,}5, 3, 3, 10, 10, 25, 25, 50, 50; 1, \overset{p-s-12)}{\ldots}, 1 \right)$.

En la Figura 4 se aprecia como hay algoritmos afectados por el fenómeno de la escala, obteniendo diferentes resultados para el caso sin estandarizar y el que aplica la estandarización univariante. En particular, el caso sin estandarización tiende a seleccionar las covariables con mayor escala, independientemente de si son relevantes o no. De nuevo, hay algoritmos que hacen uso de la estructura de los datos (en naranja), seleccionando menos de $s = 10$ covariables importantes ($FDP \downarrow$) y otros que buscan la recuperación total ($TPP \uparrow$). Solo un grupo pequeño de procedimientos se muestran "invariantes" a los efectos de la estandarización. Más detalles, así como la consideración de más escenarios de simulación, se pueden ver en el Capítulo 3.2 de Freijeiro-González (2023).

Figura 4: Número de covariables importantes (área rosa/azul oscuro) e irrelevantes (área rosa/azul claro) seleccionadas en base al caso sin estand./estand. univ. en los Escenarios 1.b (primera fila) y 1.c (segunda fila) para $n = 50$. Las líneas discontinuas marcan el valor $s = 10$.

3.3 Covariables con estructura de dependencia y en distintas escalas

En vista de los resultados arrojados en las Secciones 3.1 y 3.2, es interesante ver qué ocurre cuando se mezclan estructuras de dependencia y covariables en diferentes escalas. Un estudio de simulación y análisis de estos casos puede encontrarse en el Capítulo 3.2 de Freijeiro-González (2023).

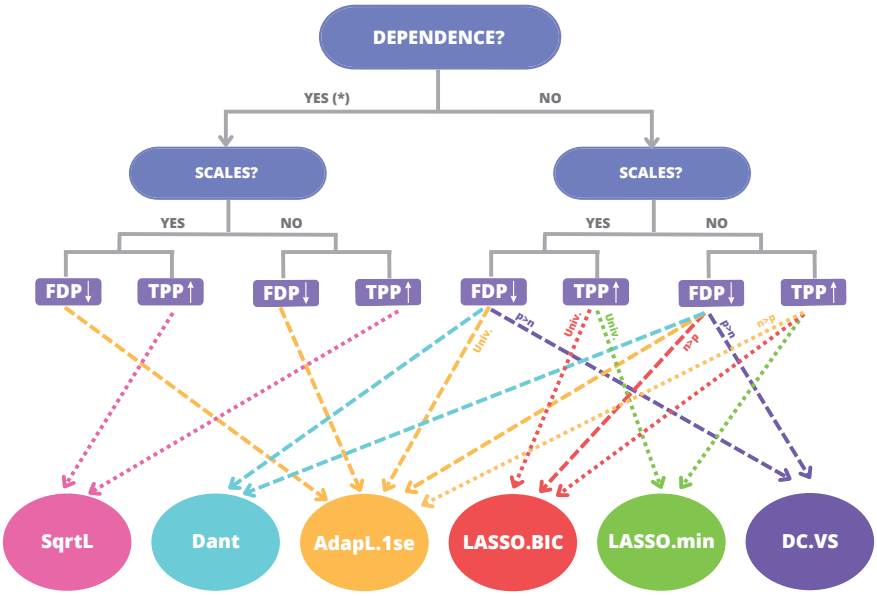


Figura 5: Mejor método de selección de covariables dependiendo de si existe alguna estructura de dependencia y/o covariables en distintas escalas. (*) indica que la opción óptima dependerá del tipo de estructura de dependencia (ver Freijeiro-González *et al.* (2022)) y $p > n$, $n > p$ o $Univ.$ (estandarización univariante) que solo aplica en esos contextos. FDP es la tasa de falsos positivos y TPP la que aplica para el número de verdaderos positivos.

Al igual que considerando únicamente el problema de la dependencia o escala, la selección de un algoritmo óptimo para seleccionar covariables dependerá del contexto en el que nos encontremos y de nuestro objetivo: $FDP \downarrow$ ó $TPP \uparrow$. En base a los resultados obtenidos en Freijeiro-González *et al.* (2022) y Freijeiro-González (2023) proponemos el árbol de decisión de la Figura 5.

4. APLICACIÓN A LOS DATOS DE RIBOFLAVINA

Una vez que hemos visto en la Sección 2 distintas opciones de procedimientos para seleccionar covariables en el contexto donde $p > n$, y que la elección de una selección adecuada dependerá de la estructura de nuestros datos (Sección 3), vamos a ver cómo implementar estas conclusiones en la práctica. Para este fin, recuperamos los datos de riboflavina introducidos en la Sección 1.

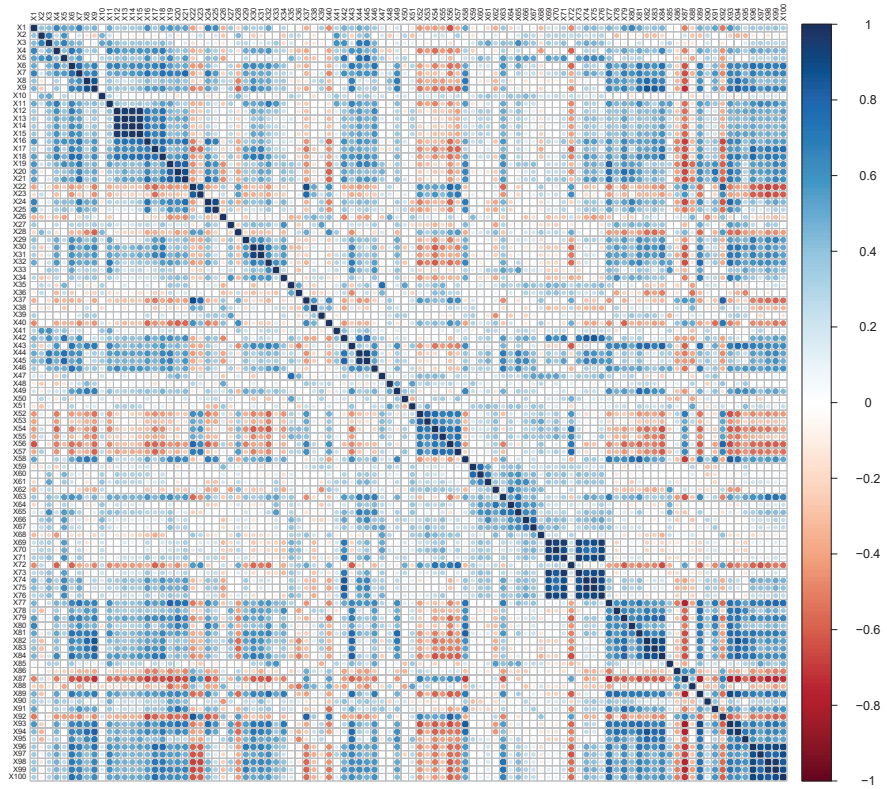


Figura 6: Matriz de correlaciones de los primeros 100 genes de los datos de riboflavina.

Para ver si existen posibles relaciones de dependencia entre los $p = 4088$ genes considerados, podemos ver cómo son las correlaciones entre ellos. Un ejemplo de la correlación entre los primeros 100 genes se muestra en la Figura 6. En esta se aprecia que existen fuertes estructuras de dependencia y que parece que todos los genes están relacionados entre sí de alguna forma. Por lo tanto, parece que con un subconjunto de los mismos debidamente seleccionado se podría explicar correctamente la producción de riboflavina. Esta situación es similar al caso de la dependencia tipo Toeplitz presentada en la Sección 3.1.

En relación a la escala de las covariables, se puede ver que estos valores se mueven en el rango $[0{,}1, 1{,}84]$. Además, se aprecian algunas diferencias entre las magnitudes de las escalas, notando que algunos de los valores más altos se alejan bastante del resto. Esta situación es similar a la presentada en la Sección 3.2. Como resultado, tiene sentido comparar los resultados trabajando sin estandarización y aplicando una estandarización univariante sobre los datos.

En consecuencia, aplicamos los procedimientos de la Sección 2 teniendo en cuenta las consideraciones de la Sección 3. La selección resultante para los casos sin estandarizar y con estandarización univariante se muestran en la Figura 7. Como era de esperar, se ven diferencias en la selección entre ambas formas de proceder, motivando la necesidad de aplicar una estandarización univariante para una correcta selección. Considerando esta, vemos que se seleccionan únicamente entre 4 y 70 genes del total de $p = 4088$, reduciendo mucho la dimensión y permitiendo la estimación del modelo.

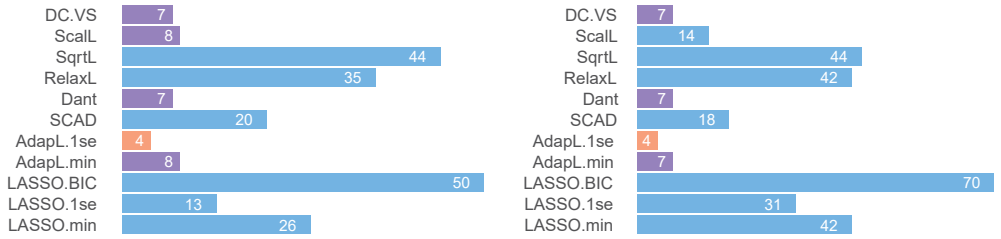


Figura 7: Número de covariables seleccionas para el caso sin estandarización (izquierda) y aplicando una estandarización univariante (derecha) en los datos de riboflavina.

En base a los resultados obtenidos en los estudios de simulación de Freijeiro-González *et al.* (2022) y en el Capítulo 3 de Freijeiro-González (2023), resumidos a lo largo de la Sección 3, se puede ver que procedimientos como el AdapL.min, AdapL.1se, Dant y DC.VS parecen los más adecuados para una selección acertada en base a $FDP \downarrow$ acorde a la naturaleza de los datos. Mientras que otras técnicas como el RelaxL, SqrtL o ScalL están enfocadas en $TPP \uparrow$. La selección génica hecha por cada algoritmo aplicando la estandarización univariante puede verse en la Tabla 1.

LASSO.min	ARGF_at, DNAJ_at, GAPB_at, LYSC_at, PRIA_at, SPOIIAA_at, SPOVAA_at, THIA_at, THIK_at, XHLB_at, XKDP_at, YACN_at, YBFI_at, YCDH_at, YCGO_at, YCKE_at, YCLB_at, YCLF_at, YDDH_at, YDDK_at, YEBC_at, YFHE_r_at, YFIO_at, YFIR_at, YHDS_r_at, YKBA_at, YKVJ_at, YLXW_at, YMFE_at, YOAB_at, YPGA_at, YQJT_at, YQJU_at, YRVJ_at, YTGB_at, YUID_at, YURQ_at, YWRO_at, YXLD_at, YXLE_at, YYBG_at, YYDA_at
LASSO.1se	ARGF_at, DNAJ_at, GAPB_at, LYSC_at, PKSA_at, SPOIISA_at, SPOVAA_at, XHLB_at, XKDS_at, XTRA_at, YBFI_at, YCDH_at, YCGO_at, YCKE_at, YCLB_at, YCLF_at, YDDH_at, YDDK_at, YEBC_at, YEZB_at,YFHE_r_at, YFIR_at, YHDS_r_at, YKBA_at, YOAB_at, YQJU_at, YRVJ_at, YURQ_at, YXLD_at, YXLE_at, YYDA_at
LASSO.BIC	ADHB_at, ALD_at, ARAA_at, ARAM_at, ARAN_at, ARGF_at, ARGH_at, DEGA_at, ECSB_at, GAPB_at, GUTR_at, LEVF_at, LYSC_at, METK_at, PHOA_at, PYRAA_at, sigM_at, SPOIVA_at, SPOVAA_at, XHLB_at, XKDB_at, XKDP_at, XLYA_at, YACN_at, YBFI_at, YBXA_at, YCLB_at, YDAO_at, YDDH_at, YDDK_at, YEBC_at, YESV_at, YETH_at, YFHE_r_at, YFIO_at, YHDS_r_at, YIST_at, YISU_at, YKBA_at, YKNV_at, YKVJ_at, YLXW_at, YMAH_i_at, YOAB_at, YOSU_at, YPGA_at, YPUI_at, YQED_at, YQGJ_at, YQJT_at, YQJU_at, YRVJ_at, YTGB_at, YTSA_at, YUID_at, YULB_at, YULC_at, YURR_at, YUSJ_at, YVFM_at, YVHJ_at, YWBI_at, YWJG_at, YWRO_at, YXAF_at, YXIB_at, YXLD_at, YXLE_at, YYBI_at, YYCO_at
AdapL.min	ARGF_at, SPOVAA_at, XHLB_at, YCLB_at, YEBC_at, YOAB_at, YXLD_at
AdapL.1se	ARGF_at, XHLB_at, YOAB_at, YXLD_at
SCAD	AADK_at, ARGC_at, IOLE_at, mrpD_at, PCP_at, SPOVAB_at, sspM_r_at, YFIJ_at, YHDT_at, YHEH_at, YKCA_at, YMFF_at, YOAC_at, YOSV_r_at, YPZE_at, YRVM_at, YXIC_at, YXLF_at
Dant	XHLA_at, XTRA_at, YCGN_at, YCKE_at, YDAR_at, YOAB_at, YXLD_at
RelaxL	ARGF_at, CTAA_at, DNAJ_at, GAPB_at, LYSC_at, PRIA_at, SPOIIAA_at, SPOVAA_at, THIA_at, THIK_at, XHLB_at, XKDB_at, YACN_at, YBFI_at, YCKE_at, YCLB_at, YCLF_at, YDDH_at, YDDK_at, YEBC_at, YFHE_r_at, YFIO_at, YFIR_at, YHDS_r_at, YKBA_at, YKVJ_at, YLXW_at, YMFE_at, YOAB_at, YPGA_at, YQJT_at, YQJU_at, YRVJ_at, YTGB_at, YUID_at, YWRO_at, YXIB_at, YXLD_at, YXLE_at, YYBG_at, YYCO_at, YYDA_at
SqrtL	LYSC_at, METB_at, PHRI_r_at, RPLJ_at, RPLL_at, RPLO_at, RPLP_at, RPLX_at, RPSN_at, SIGY_at, XHLA_at, XKDS_at, XTRA_at, YBGB_at, YCDH_at, YCDI_at, YCEA_at, YCGM_at, YCGN_at, YCGO_at, YCGP_at, YCKE_at, YCLF_at, YDAR_at, YDBM_at, YDDK_at, YDDM_at, YEBC_at, YHFH_r_at, YHZA_at, YOAB_at, YODF_at, YRPE_at, YRVJ_at, YTGA_at, YTGB_at, YTGD_at, YTIA_at, YXLC_at, YXLD_at, YXLE_at, YXLF_at, YXLG_at, YXLJ_at
ScalL	LYSC_at, SPOIISA_at, XHLA_at, XKDS_at, XTRA_at, YCGN_at, YCGO_at, YCKE_at, YDDK_at, YEBC_at, YHCL_at, YOAB_at, YURQ_at, YXLD_at
DC.VS	FLHO_at, RPLX_at, xepA_at, YCKE_at, YQKD_at, YRHC_at, YWRO_at

Tabla 1: Genes seleccionados en los datos de riboflavina para el caso de estandarización univariante 9/11 (**coral**), 7/11 (**violet**) y 6/11 (**blue**) veces.

Bühlmann *et al.* (2014) han detectado tres genes como relevantes: `LYSC_at`, `YOAB_at` y `YXLD_at`. Vemos en la Tabla 1 que estos dos últimos se detectan por 9 de los 11 algoritmos considerados, mientras que `LYSC_at` por 6 procedimientos. La selección más parecida sería la realizada por el AdapL.1se que incluye los dos últimos y añade además `ARGF_at` y `XHLB_at`. Puesto que el AdapL.1se se basa en $FDP \downarrow$ y ha demostrado incluir muy poco o nada de ruido, se tendrían que considerar los genes `ARGF_at` y `XHLB_at` como posibles covariables importantes. Así mismo, la no inclusión de `LYSC_at` se podría explicar por la necesidad de un tamaño muestral mayor o porque este último está enmascarando a otros genes relevantes.

5. CONCLUSIONES

En estudios genéticos es muy común encontrarse con el hecho de que se dispone de un mayor número de genes a analizar que de muestras ($p > n$). En consecuencia, como se explicó en la Sección 1, tiene especial interés realizar un procedimiento de selección de covariables para considerar únicamente aquellos genes que proporcionan información relevante, aliviando los problemas de la alta dimensión. Sin embargo, los métodos usuales no sirven en este contexto y es necesario recurrir a procedimientos diseñados para el mismo (Sección 2). En la literatura es muy común aplicar estos sin considerar los tipos de estructura de dependencia que existen en los datos, así como sin analizar el efecto de las escalas de las covariables. En la Sección 3 se resumen los resultados arrojados mediante diferentes estudios intensivos de simulación para estas dos cuestiones. Como conclusión, dependerá del tipo de estructura de dependencia con la que estemos trabajando, de nuestra finalidad y de si las covariables se asumen o no en la misma escala. De esta forma, una mala selección puede llevar a resultados erróneos o incompletos. A modo de guía, proponemos un árbol de decisión para seleccionar la metodología más adecuada para cada caso en la Figura 5.

Una buena selección genética permite entender mejor qué genes influyen en ciertos procesos fisiológicos, qué genes son los que afectan en la predilección a sufrir una determinada enfermedad o incluso pueden ayudar a desarrollar tratamientos personalizados para pacientes con enfermedades raras donde el tamaño muestral es limitado. Un ejemplo de cómo proceder en la práctica se ilustra a través de los datos de la producción de riboflavina de la bacteria *Bacillus subtilis* (Bühlmann *et al.* (2014)). En la Sección 4 aplicamos las conclusiones de la Sección 3 a esta base de datos y comparamos nuestros resultados con los obtenidos por Bühlmann *et al.* (2014), descubriendo nuevos genes que pueden ser importantes en la producción de riboflavina.

A lo largo de este trabajo nos hemos centrado en la formulación lineal del modelo de regresión (1), sin embargo, se pueden extender las ideas de los métodos de penalización (Sección 2.1) a estructuras más complejas. Véase por ejemplo Ravikumar *et al.* (2009), Vidaurre *et al.* (2013) o Haris *et al.* (2022). Así mismo, existen extensiones de la DC (Sección 2.2) para testar no solo independencia, sino otros tipos de relaciones como independencia parcial (Székely y Rizzo (2014)), condicional en media (Shao y Zhang (2014)) o condicional (Wang *et al.* (2015)).

AGRADECIMIENTOS

Esta investigación está financiada por la Consellería de Cultura, Educación e Ordenación Universitaria junto con la Consellería de Economía, Emprego e Industria de la Xunta de Galicia (proyecto ED481A-2018/264). Así mismo, por el Proyecto PID2020-116587GB-I00 fundado por MCIN/AEI/10.13039/501100011033, por "ERDF A way of making Europe", junto con los Grupos Competitivos de Referencia 2021-2024 (ED431C 2021/24) de la the Xunta de Galicia mediante el Fondo Europeo de Desarrollo Regional (ERDF). También se agradece al Centro de Supercomputación de Galicia (CESGA) por los recursos computacionales prestados.

REFERENCIAS

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. En *Selected papers of Hirotugu Akaike*, pp. 199–213. Springer.

Belloni, A., Chernozhukov, V., y Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Bühlmann, P. y Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1 + l2 penalization. *Electronic Journal of Statistics*, 2:1153–1194.

Bühlmann, P., Kalisch, M., y Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278.

Candes, E. y Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Fan, J. (1997). Comments on «wavelets in statistics: A review» by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131.

Febrero-Bande, M., González-Manteiga, W., y Oviedo de la Fuente, M. (2019). Variable selection in functional additive regression models. *Computational Statistics*, 34(2):469–487.

Freijeiro-González, L. (2023). *New covariates selection approaches in high dimensional or functional regression models*. Tesis doctoral, Universidade de Santiago de Compostela (España): http://hdl.handle.net/10347/30892.

Freijeiro-González, L., Febrero-Bande, M., y González-Manteiga, W. (2022). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *International Statistical Review*, 90(1):118–145.

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC.

Giraud, Christophe and Huet, Sylvie and Verzelen, Nicolas and others (2012). High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518.

Haris, A., Simon, N., y Shojaie, A. (2022). Generalized sparse additive models. *Journal of Machine Learning Research*, 23(70):1–56.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., y Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

Hoerl, A. E. y Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Homrighausen, D. y McDonald, D. J. (2018). A study on tuning parameter selection for the high-dimensional lasso. *Journal of Statistical Computation and Simulation*, 88(15):2865–2892.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1):25–45.

Ravikumar, P., Lafferty, J., Liu, H., y Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, X. y Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

Su, W., Bogdan, M., y Candès, E. (2017). False discoveries occur early on the Lasso path. *The Annals of statistics*, 45(5):2133–2150.

Sun, T. y Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Székely, G. J. y Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382 – 2412.

Szekely, G. J. y Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4:447–479.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the LASSO: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Vidaurre, D., Bielza, C., y Larranaga, P. (2013). A survey of l1 regression. *International Statistical Review*, 81(3):361–387.

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

Wasserman, L. y Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201.

Wissler, C. (1905). The spearman correlation formula. *Science*, 22(558):309–311.

Yuan, M. y Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhao, P. y Yu, B. (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

FORMULACIÓN DEL PROBLEMA		PROS
LASSO – Tibshirani (1996) $\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert \right\}$	✓	–
▲ **SCAD** – Fan (1997) $\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + p_\lambda(\beta) \right\}$ con $p_\lambda(\beta) = \begin{cases} \lambda\lvert\beta\rvert, & \text{si } \lvert\beta\rvert \le \lambda, \\ \dfrac{2a\lambda\lvert\beta\rvert - \beta^2 - \lambda^2}{2(a-1)}, & \text{si } \lambda < \lvert\beta\rvert \le a\lambda \quad (a > 2) \\ \dfrac{\lambda^2(a+1)}{2}, & \text{otro.} \end{cases}$	×	*Mejor selección* *Reducción del sesgo*
● **Adaptive LASSO** – Zou (2006) $\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} w_j \lvert\beta_j\rvert \right\}$ (tomando $w_j = 1/\lvert\hat{\beta}_j^{RIDGE}\rvert^q$ donde $\hat{\beta}^{RIDGE}$ es el estimador RIDGE (Hoerl y Kennard (1970)) y $q \ge 1$)	✓	*Mejor selección* *Reducción del sesgo*
▲ **Dantzig selector** – Candes y Tao (2007) $\min_{\beta} \lVert\beta\rVert_1 \quad \text{s. a. } \lVert X^\top r\rVert_\infty \le \lambda_p \cdot \sigma$ (con $\lVert X^\top r\rVert_\infty := \sup_{1 \le j \le p} \lvert(X^\top r)_j\rvert$ y $r = y - X\beta$)	×	*Consistente a transformaciones ortogonales*
▲ **Relaxed LASSO** – Meinshausen (2007) $\min_{\beta} \left\{ n^{-1} \sum_{i=1}^{n} \left(y_i - x_i^\top \{\beta \cdot \mathbf{1}_{\mathcal{M}_\lambda}\}\right)^2 + \theta\lambda\lVert\beta\rVert_1 \right\} \text{ con } \theta \in (0, 1]$	✓	*Tasas rápidas de convergencia* *Predicciones más precisas*
▲ **Square-root LASSO** – Belloni *et al.* (2011) $\min_{\beta} \left\{ \left[\sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 \right]^{1/2} + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert \right\}$	✓	*No se necesita conocer σ para obtener el λ óptimo*
▲ **Scaled LASSO** – Sun y Zhang (2012) $\hat{\sigma} \leftarrow \lVert y - X^\top \hat{\beta}^{old}\rVert_2 / n^{1/2}, \quad \lambda \leftarrow \hat{\sigma}\lambda_0$ $\hat{\beta}^{new} = \arg\min_{\beta} \begin{cases} x_j^\top (y - X^\top \hat{\beta})/n = \lambda \text{sign}(\hat{\beta}_j), & \text{si } \hat{\beta}_j \ne 0, \\ x_j^\top (y - X^\top \hat{\beta})/n \in \lambda[-1, 1], & \text{si } \hat{\beta}_j = 0. \end{cases}$ $\hat{\beta} \leftarrow \hat{\beta}^{new}, \; L_\lambda(\hat{\beta}^{new}) \le L_\lambda(\hat{\beta}^{old})$ $\left(\text{donde } L_\lambda(\beta) = \frac{\lVert y - X^\top \beta\rVert_2^2}{2n} + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert\right)$	✓	*Estimación simultánea de σ y β*

Tabla 2: Formulación de problemas para la estimación del vector β en la regresión lineal en un contexto de alta dimensión ($p > n$). Se indica si los problemas de optimización son convexos (✓) o no (×). Sus principales ventajas en comparación con el LASSO se muestran en la columna **PROS** y se indica si son versiones ponderadas del LASSO (●) o alternativas (▲).

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Diagramas Causales: Herramienta Innovadora para Comprender y Mejorar la Práctica Clínica en Residentes de Enfermería y Medicina

Sara Rodríguez Pastoriza[1], Macarena Chacón Docampo[1,2], David Liñares Mariñas[1], Ángela Asensio Martínez[3], Ana Clavería Fontán[1,2], María Victoria Martín Miguel [1,2,4], Clara González Formoso[2,4] y Javier Roca Pardiñas[2,5,6].

[1] Instituto de Investigación Sanitaria Galicia Sur, Área de Salud de Vigo, Servicio Galego de Saúde, Vigo, España.
[2] Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud/RICAPPS, Vigo, España.
[3] Área de Psicología Social de la Universidad de Zaragoza, Zaragoza, España.
[4] Unidad Docente Multiprofesional de Medicina Familiar y Comunitaria de Vigo y Enfermería, Área de Salud de Vigo, Servicio Galego de Saúde, Vigo, España.
[5] Departamento de Estadística e Investigación Operativa, Universidad de Vigo, Vigo, España.
[6] Centro de Investigación y Tecnología Matemática de Galicia (Centro de Investigación e Tecnoloxía Matemática de Galicia/CITMAga), Santiago de Compostela, España.

RESUMEN

El Burnout es un tipo de agotamiento físico y mental que se caracteriza por una sensación de agotamiento extremo, desagregación hacia el trabajo y disminución de la eficacia laboral. Los profesionales de la salud tienden a sufrir este síndrome con mayor frecuencia, principalmente debido a la toma de decisiones que conlleva su trabajo diario. Se sabe que los principales factores que influyen en este síndrome son estresores ajenos al individuo como, por ejemplo, la carga de trabajo excesiva, horarios de trabajo inmanejables, carga administrativa elevada o angustia moral. Por todo lo anterior, el propósito de este estudio es analizar las relaciones causales que se establecen entre los diferentes aspectos sociodemográficos, laborales y psicológicos de médicos y enfermeros residentes, y en qué medida pueden afectar estas relaciones al agotamiento profesional.

Palabras y frases clave: Diagramas causales, variables mediadoras, relaciones causales, Burnout, residentes, familiar y comunitaria.

1. INTRODUCCIÓN

El Burnout fue descrito por primera vez en 1974 por Freudenberger como "la sensación de fracaso y una existencia agotada o gastada que resulta de una sobrecarga por exigencias de energías, recursos personales o fuerza espiritual del trabajador"[1]. Sin embargo, es en el año 2022 cuando se incluye en la Clasificación Internacional de Enfermedades (CIE-11) dentro del capítulo Factores que influyen en el estado de salud o el contacto con los servicios de salud, dentro de la subcategoría de problemas asociados con el empleo y el desempleo y codificado como "Síndrome de desgaste ocupacional", con el código QD85.

La National Academy of Medicine pone de relieve este problema, destacando que los sanitarios con frecuencia presentan prevalencias de Burnout superiores al 50%, que se acompaña de consecuencias perjudiciales tanto para el individuo (abuso sustancias, depresión, suicidio...) como para la sociedad. Por otra parte, una de las consecuencias más respaldada por la bibliografía es el empobrecimiento de la alianza terapéutica, desencadenada por una pérdida de empatía del clínico hacia el paciente a causa del agotamiento emocional que sufre [2]. Los principales factores que influyen en este síndrome son estresores ajenos al individuo y relacionados con la organización del

trabajo: carga de trabajo excesiva, horarios de trabajo inmanejables, personal inadecuado, carga administrativa elevada, interrupciones y distracciones, inadecuada usabilidad tecnológica, precisión del tiempo y usurpación del tiempo personal, angustia moral, factores del paciente, etc. A primera vista, las estrategias para abordar este problema son las centradas en la organización del trabajo, pero tendrían que ser realizadas por los organismos políticos o de gestión competentes. Por eso, las habilidades centradas en la persona, a menudo, son las únicas herramientas que se tienen para abordar el problema [2] en equipos clínicos o docentes.

Por todo lo anterior, nos planteamos visualizar y analizar las relaciones causales entre las diferentes variables que puedan influir en el síndrome del Burnout personal, del Burnout de trabajo y del Burnout de trabajo con clientes a través de diagramas causales. Estas herramientas estadísticas son especialmente útiles cuando se tiene un conjunto de variables interconectadas y se desea comprender cómo se relacionan entre sí y cómo influyen unas en otras [3].

El propósito de este trabajo es, por tanto, analizar las relaciones causales entre las diferentes variables contextuales que pueden influir en el síndrome del Burnout en residentes de las Unidades Docentes Multiprofesionales de Atención Familiar y Comunitaria de Vigo, Zaragoza y Mallorca.

2. METODOLOGÍA

Diseño del estudio

Estudio descriptivo de causalidad. El proyecto ha sido registrado en Clinical Trials siendo su ID: NCT04625582. Aprobado por el Comité de Ética e Investigación Clínica de Vigo.

Objetivo principal

Analizar las relaciones causales que puedan influir en el nivel de Burnout personal, Burnout de trabajo y Burnout de trabajo con pacientes en residentes de las Unidades Docentes Multiprofesionales de Familiar Y Comunitaria (UDMFYC).

Ámbito y población

Residentes de enfermería y medicina de Atención Familiar y Comunitaria de las promociones 2018, 2019, 2020 y 2021 de las Unidades Docentes de Vigo, Zaragoza y Mallorca (N= 340).

Análisis estadístico

Tras depurar los datos y seguir el protocolo habitual, se realizó un análisis exploratorio inicial. La normalidad de las variables cuantitativas se estudió con la prueba de Shapiro Wilk considerando una significación nivel $\alpha = 0.05$. Se realizó el análisis descriptivo considerando la media aritmética y la desviación estándar para variables cuantitativas, y la frecuencia y porcentaje para variables cualitativas. Se llevó a cabo un análisis bivariante según el sexo para determinar si existían o no diferencias significativas entre los dos grupos. Se aplicó la prueba de chi-cuadrado para las variables cualitativas y la prueba U de Mann-Whitney para las variables cuantitativas. En el análisis de la causalidad, se llevó a cabo el cálculo de las correlaciones de pearson entre todas las variables de estudio, a partir de las cuales se diseñaron posteriormente los diagramas causales. Una vez diseñados los diagramas causales, se llevó a a cabo el análisis de las relaciones causales haciendo uso de la función *sem()* del paquete estadístico lavaan [4].

Variables e instrumentos

Síndrome de Burnout (medido con el cuestionario validado Copenhagen Burnout Inventory).

Sociodemográficas: edad, sexo, nacionalidad, profesión, promoción, Comunidad Autónoma.

Laborales: promedio de horas trabajadas por semana, promedio mensual de llamadas, número de pacientes por día en la consulta, especialidad.

Aspectos psicológicos: empatía (Índice de Reactividad), resiliencia (10 ítems CD-RISC), apoyo social (OSLO-3), sentido de coherencia (OLQ-13), ansiedad y depresión (HADS), rasgos de personalidad (TIPI).

Resultados

De los 340 residentes de las diferentes Unidades Docentes Multiprofesionales de Familiar Y Comunitaria (UDMFYC), participaron un total de 177 residentes. El análisis descriptivo de las variables sociodemográficas cualitativas de referencia mostró que: el 82% de la muestra eran mujeres, el 69,9% eran médicos, el 42,6% realizaban la residencia en la comunidad autónoma de Galicia, el 40,3% en la comunidad autónoma de Aragón y el 17,1% en la comunidad autónoma de Baleares. El valor medio de la edad fue de 27,69 (DE: 3,86), de las horas de trabajo a la semana fue de 49,16 (DE: 10,21), de los pacientes por consulta al día de 30,46 (DE: 14,05) y de las guardias al mes de 4,40 (DE: 1,45) (Tabla 1). Respecto al análisis descriptivo de los aspectos psicológicos (Tabla 2) se muestra la media (DE) y los valores perdidos de cada variable, junto con el rango real y el rango posible.

	Sociodemográficas-laborales[#]	
	Total	**NA**
N	177	
Sexo (mujer)	141 (82,0)	5
Edad	27,69 (3,46)	8
Especialidad (medicina)	123 (69,9)	1
Lugar de residencia		1
Aragón	71 (40,3)	
Galicia	75 (42,6)	
Baleares	30 (17,1)	
Promoción		5
2017	2 (1,2)	
2018	33 (19,2)	
2019	38 (22,1)	
2020	55 (32,0)	
2021	44 (25,6)	
Horas de trabajo (semana)	49,16 (10,21)	3
Pacientes en consulta (día)	30,46 (14,05)	7
Guardias (mes)	4,40 (1,45)	1

[#] Los datos se expresan en media (desviación estándar) para cuantitativas y en número (%) para cualitativas.

Tabla 1: Análisis descriptivo de las variables sociodemográficas-laborales.

Aspectos psicológicos[#]				
	Media (DE)	**Rango real**	**Rango posible**	**NA**
Burnout (CBI)				
Personal	45,75 (17,97)	0,00-95,83	0,00-100,00	
Trabajo	47,34 (16,83)	0,00-95,83	0,00-100,00	3
Trabajo con pacientes	37,90(15,91)	0,00-95,83	0,00-100,00	
Empatía (IRI)				4
Fantasía	22,45 (5,59)	9,00-34,00	0,00-35,00	
Perspectiva	26,02 (4,09)	14,00-35,00	0,00-35,00	
Preocupación empática	25,71 (3,17)	17,00-33,00	0,00-35,00	
Angustia personal	14,09 (4,21)	7,00-27,00	0,00-35,00	
Resiliencia (CD-risk)	26,58 (5,55)	12,00-39,00	0,00-40,00	4
Ansiedad (HADS)	7,36 (3,51)	1,00-17,00	0,00-21,00	4
Depresión (HADS)	4,65 (2,76)	1,00-12,00	0,00-21,00	4
Apoyo social (OSLO)	10,60 (1,60)	4,00-13,00	3,00-14,00	4
Sentido de coherencia (SOC)	63,07 (11,49)	29,00-86,00	13,00-91,00	5
Personalidad (TIPI)				4
Extraversión	4,56 (1,18)	1,50-7,00	1,00-7,00	
Afabilidad	4,51 (1,15)	1,50-7,00	1,00-7,00	
Responsabilidad	4,07 (0,95)	1,50-7,00	1,00-7,00	
Estabilidad emocional	4,22 (1,02)	1,50-7,00	1,00-7,00	
Apertura	6,11 (0,74)	3,50-7,00	1,00-7,00	

[#] Los datos como mínimo-máximo para los rangos.

Tabla 2: Análisis descriptivo de los aspectos psicológicos.

En el análisis bivariante por sexo, solo mostró diferencias significativas la variable independiente responsabilidad (TIPI) en el factor de responsabilidad, con un intervalo de confianza al 95% de (4,03 - 4,41) en mujeres y (3,25 - 3,95) en hombres (p valor: 0,003).

Sociodemográficas-laborales[#]			
	Femenino	**Masculino**	**p valor**
N	141 (82,0)	31 (18,0)	
Edad	27,7 (27,1 - 28,3)	27,8 (26,7 - 29,0)	0,780
Especialidad			0,078
Medicina	65,7 (57,2 - 73,5)	83,9 (66,3 - 94,5)	
Enfermería	34,3 (26,5 - 42,8)	16,1 (5,45 - 33,7)	
Lugar de residencia			0,992
Aragón	42,9 (34,5 - 51,5)	41,9 (24,5 - 60,9)	
Galicia	40,7 (32,5 - 49,3)	41,9 (24,5 - 60,9)	
Baleares	16,4 (10,7 - 23,6)	16,1 (5,45 - 33,7)	
Promoción			0,699
2017	1,42 (0,17 - 5,03)	0,00 (0,00 - 11,2)	
2018	17,7 (11,8 - 25,1)	25,8 (11,9 - 44,6)	
2019	21,3 (14,8 - 29,0)	25,8 (11,9 - 44,6)	
2020	33,3 (25,6 - 41,8)	25,8 (11,9 - 44,6)	
2021	26,2 (19,2 - 34,3)	22,6 (9,59 - 41,1)	
Meses en el centro de salud	2,69 (2,40 - 2,97)	3,08 (2,41 - 3,75)	0,279
Horas de trabajo (semana)	48,1 (46,3 - 49,9)	52,1 (47,9 - 56,2)	0,079
Pacientes en consulta	30,3 (28,0 - 32,7)	31,4 (25,7 - 37,1)	0,725
Guardias (mes)	5,74 (4,62 - 6,85)	4,87 (4,52 - 5,22)	0,143

[#] Los datos se expresan en media (IC 95%) para cuantitativas y en porcentaje (IC 95%) para cualitativas.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

	Aspectos psicológicos[#]		
	Femenino	**Masculino**	**p valor**
Burnout (CBI)			
Personal	45,9 (43,3 - 48,6)	43,1 (35,2 - 51,1)	0,504
Trabajo	47,3 (44,8 - 49,9)	50,5 (43,5 - 57,6)	0,388
Trabajo con pacientes	37,3 (35,0 - 39,6)	45,0 (38,0 - 52,1)	0,041
Empatía (IRI)			
Fantasía	22,4 (21,4 - 23,3)	22,6 (20,3 - 24,9)	0,834
Perspectiva	25,9 (25,2 - 26,6)	26,5 (25,0 - 27,9)	0,444
Preocupación empática	25,9 (25,3 - 26,4)	24,8 (23,8 - 25,9)	0,076
Angustia personal	14,2 (13,5 - 14,9)	13,5 (12,2 - 14,9)	0,379
Resiliencia (CD-risk)	26,3 (25,3 - 27,2)	27,6 (25,6 - 29,6)	0,217
Ansiedad (HADS)	7,39 (6,81 - 7,97)	7,06 (5,77 - 8,36)	0,641
Depresión total (HADS)	4,67 (4,21 - 5,14)	4,52 (3,54 - 5,50)	0,772
Apoyo social (OSLO)	10,6 (10,4 - 10,9)	10,5 (9,87 - 11,2)	0,775
Sentido de coherencia (SOC)	62,9 (61,0 - 64,8)	63,8 (59,5 - 68,1)	0,712
Personalidad (TIPI)			
Extraversión	4,55 (4,35 - 4,75)	4,65 (4,23 - 5,06)	0,668
Afabilidad	4,44 (4,25 - 4,63)	4,85 (4,47 - 5,24)	0,058
Responsabilidad	4,19 (4,03 - 4,34)	3,60 (3,25 - 3,95)	0,003**
Estabilidad emocional	4,25 (4,09 - 4,41)	4,02 (3,55 - 4,48)	0,337
Apertura	6,14 (6,03 - 6,25)	5,90 (5,53 - 6,27)	0,216

[#] Los datos se expresan en media (IC 95%) para cuantitativas y en porcentaje (IC 95%) para cualitativas.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Entre los resultados del análisis de la correlación destaca la especialidad, que se correlacionó positivamente y de manera significativa con las horas de trabajo (0,39; p<0,001), los pacientes al día en consulta (0,57; p<0,001) , las guardias al mes (0,37; p<0,001) y el Burnout de trabajo con pacientes (0,28; p<0,001). Sin embargo, correlacionó negativamente y de manera significativa con la promoción (-0,39; p<0,001). El Burnout personal, de trabajo y de trabajo con pacientes correlacionaron positivamente y de manera significativa, con una correlación de Pearson superior a 0,4. El sentido de coherencia correlacionó negativamente y de manera significativa con el Burnout personal (-0,52; p<0,001) , con el Burnout de trabajo (-0,54; p<0,001), Burnout trabajo con pacientes (-0,47; p<0,001), con la ansiedad (-0,71; p<0,001) y con la depresión (-0,57; p<0,001). En contraste, correlacionó positivamente y de manera significativa con resiliencia (0,54; p<0,001) y el apoyo social (0,33; p<0,001). Los restantes coeficientes de correlación de Pearson y significancias no resultaron relevantes.

A partir de los resultados de coeficientes de correlación obtenidos, se presentarán los análisis de los diagramas causales que resultaron de interés.

Discusión y conclusiones

Se presentarán en la comunicación correspondiente.

Bibliografía

[1] Freudenberger, H. J. Staff Burnout. J Soc Issues. 1974;30(1):159-165.

[2] Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being, National Academy of Medicine, National Academies of Sciences, Engineering, and Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being [Internet]. Washington, D.C.: National Academies Press; 2019 [citado 8 de noviembre de 2022]. Disponible en: https://www.nap.edu/catalog/25521

[3] Cunningham, S. D. (2021). Causal Inference: The Mixtape. Yale University Press.

[4] Rosseel Y. lavaan: An R Package for Structural Equation Modeling [Internet]. Vol. 48, Journal of Statistical Software. 2012. p. 1–36. Available from: http://www.jstatsoft.org/v48/i02/

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26, 27 e 28 de outubro de 2023

DUNHA NOVA TABULADORA BASEADA EN VARIACIÓNS RELATIVAS

Carlos L. Iglesias Patiño[1]

[1] Instituto Galego de Estatística

RESUMO

Velaquí unha reflexión sobre a variación relativa entre dous valores dunha variábel positiva estudando diferentes fórmulas cun duplo fin: divulgativo e innovador. Tras unha breve formalización unha delas é escollida, a diferenza de logaritmos, para favorecer a súa interpretación des dun punto de vista descritivo e xeneralizar a súa utilidade na construción dunha nova tabuladora, baseada no criterio dos cadrados mínimos e nesta fórmula ubicua na estatística.

Palabras e frases chave: tabuladora, cocientes, suavización local, puntos porcentuais

1. INTRODUCIÓN

Existe a necesidade de alisar ou suavizar cocientes: taxas, fraccións ou razóns. Nestes casos pode que o mellor había ser empregar a razón local (RaLo), a R_1 en Iglesias (2019), mais non sempre están dispoñíbeis os macro ou meso-datos precisos.

Por outra banda, na estatística pública ou nas ciencias sociais cuantitativas, os criterios de optimalidade ou erro baseados nas diferenzas absolutas poden non ser axeitados e convén consideralas en termos relativos, por exemplo (p.ex.) para independizalas das unidades de medida.

A case omnipresencia da variación relativa en termos percentuais no eido socioeconómico, herdado da aritmética mercantil, é comprensíbel no estudo de variábeis referidas a distintos intres ou lapsos temporais. Porén no estudo transversal, sincrónico, pode resultar menos claro e mesmo carecer de sentido na comparanza de cocientes.

Todos temos oído falar da diferenza en "puntos porcentuais" ou topado con compararmos taxas, verdadeiras ou simples fraccións ou razóns. Tanto a demografía como a econometría teñen avanzado neste asunto. A última, no estudo lonxitudinal ou diacrónico, aproximándose ás outras ciencias e técnicas ao transformar mediante logaritmos as series temporais. E a primeira cun tratamento específico das taxas, diferenciándoas dos cocientes "prospectivos" (probabilidades) e retrospectivos.

2. IMPARIDADES

Para comparar dous valores dunha variábel, empregamos a diferenza $\Delta(y,x) = y - x$ ou incremento de x a y. Esta permite caracterizar a tricotomía $y \lessgtr x \Leftrightarrow \Delta(y,x) \lessgtr 0$. Cando x é un valor obxectivo e y unha estimativa del, $\Delta(y,x)$ é o erro que cometemos ao empregarmos y no canto de x. De prescindirmos do seu sinal (signo), estamos perante o erro absoluto. Este último é empregado como perda, elemental, na función de perdas ou ben o seu cadrado. O primeiro, a diferenza, tamén se emprega noutros eidos da estatística como p.ex. no numerador das ratios de Student para contrastes de hipóteses.

O problema da diferenza é que depende das unidades e non se comporta axeitadamente con diferentes ordes de grandor da variábel. Unha maneira de resolver isto é considerarmos o incremento relativo $\Delta(y,x)/x$ semellante ao que acontece co tipo de xuro ou rédito. De considerarmos o valor absoluto del, estamos perante o erro relativo. Nestes dous casos o x e o y non desempeñan o mesmo papel, ben por ser o obxectivo –para un exemplo en contabilidade previsional, Caamaño (2008)– ou ben pola seta do tempo. Neste último caso adoita denominarse taxa de variación relativa. Habemos denotar o incremento relativo por TVR inda que non teña compoñente temporal.

Como exemplos temos que a variancia relativa 'rel-variance' pode verse como a TVR do momento de segundo orde a respecto do cadrado do momento de primeiro. Para aplicarmos isto ao indicador HH, consulte Iglesias (2021),

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

Nesta referencia tamén podemos ver a co-primacía como o valor absoluto da TVR, sempre negativa, da norma-∞ a respecto da norma-1, agás máxima concentración.

Cando non podemos singularizar o x fronte ao y, esta solución non é satisfactoria. Mesmo cando é posible, non se adopta como no caso da converxencia σ (Cuadrado et al., 1998). Para procurar outra, imos introducir primeiro unha función $\eth(\cdot,\cdot)$ que habemos denominar imparidade mentres non houber mellor proposta, que for:

- característica da identidade, $\eth(y,x) = 0 \Rightarrow x = y$,
- indicatriz da orde ou conforme co sinal, $y \lesseqgtr x \Rightarrow \eth(y,x) \lesseqgtr 0$,
- alternada, $\eth(y,x) = -\eth(x,y)$ e
- homoxénea de grao 0

Por ela ser alternada, $x = y \Rightarrow \eth(x,x) = 0$. Por ser 0-homoxénea, non depende de factores de escala.

Os tres primeiros puntos caracterizan unha especie de distancia orientada (con sinal), que non métrica, e o cuarto, relativa. Un exemplo é a taxa de variación demográfica (TVD) $\eth(y,x) = (y-x)/((x+y)/2)$, non así o incremento relativo, $\Delta(y,x)/x$, que non é alternada. Alén disto, a TVD ten menos problemas de definición.

No entanto, sen entrarmos nestes problemas, podemos definir outra imparidade para variábeis positivas, a diferenza ou incremento na escala logarítmica, $\eth(y,\mathrm{x}) = \log y - \log x$ que habemos denominar diferenza logarítmica e denotar DOL porque pode verse como a distancia orientada logarítmica e reservarmos taxa de variación logarítmica para as situacións temporais. Se o logaritmo é neperiano poderiamos empregar taxa de variación natural seguindo a Uriel y Gea (1997). Tamén pode consultar esta referencia para ver que esta imparidade semella o incremento relativo para valores x e y próximos.

Os estatísticos de asociación ou contraste para unha táboa das continxencias poden verse como combinacións cónicas de discrepancias entre recontos observados e esperados. No de razón de verosemellanzas aparece $\eth(O_c, E_c) = \ln O_c - \ln E_c$, verdadeira imparidade, mentres que no ghi-cadrado aparece o incremento relativo $\Delta(O_c, E_c)/E_c$ (Agresti, 2002)

$$2\sum_{c=1}^{C} O_c \ln\frac{O_c}{E_c}$$

$$\sum_{c=1}^{C}\left(\frac{O_c - E_c}{E_c}\right)^2 E_c$$

O valor de referencia é o esperado en ambos os dous, o que comporta menos problemas de definición, mais os coeficientes da combinación linear conmutan, agás constante, nun caso son os esperados e noutro os observados. En xeral, o estatístico do contraste da razón de verosemellanzas, o logaritmo dela agás constante, pode interpretarse como a imparidade entre as verosemellanzas nas dúas hipóteses

3. CONSTRUCIÓN

Imos introducir, a seguir, unha nova tabuladora (Iglesias, 2021), a tabuladoX, baseada no criterio de cadrados mínimos en termos relativos; isto é (i.e.), no canto de empregar a diferencia entre y e $\hat{y}$ poderiamos pensar en empregar o incremento relativo, no entanto habemos empregar a súa imparidade $\eth(\hat{y}_i, y_i)$ con $\hat{y}_i = Rx_i$. Para o caso da DOL

$$\sum(\log y_i - \log Rx_i)^2 = \sum_{i=1}^{I}\left(\log\frac{y_i}{Rx_i}\right)^2 = \sum_{i=1}^{I}\left(\log\frac{y_i/x_i}{R}\right)^2 = \sum_{i=1}^{I}\left(\log\frac{y_i}{x_i} - \log R\right)^2$$

A primeira igualdade amosa que os "residuos" son invariantes a cambios de escala. A segunda, exprésaa como log-ratio, tan familiar no caso de fraccións na análise de composicións, Aitchison (1982). Outras propiedades e aplicacións deste criterio noutros contextos xa foron presentadas en Vilar et al. (2009) e en Prasada Rao (2001).

Por conseguinte, a tabuladoX será o resultado do seguinte problema de minimización

$$R(c) = R(B_c) = \arg\min_R \sum_i \left(\log\frac{y_i}{Rx_i}\right)^2 1_{B_c}(i) = \arg\min_R \sum(\log y_i - \log Rx_i)^2\, 1_{B_c}(i)$$

$$= \arg\min_R \sum(\log y_i - \log R - \log x_i)^2\, 1_{B_c}(i)$$

onde B_c é unha clase da partición do colectivo (Iglesias, 2021). Derivándomos a respecto de R e igualándomos a cero

$$0 = -2\sum_i \left(\log\frac{y_i}{x_i} - \log R\right)\frac{1}{R}1_{B_c}(i) = \frac{-2}{R}\sum_i \left(\log\frac{y_i}{x_i} - \log R\right)1_{B_c}(i),$$

e simplificándomos, obtemos que

$$\sum_{B_c} \log\frac{y_j}{x_j} = \sum_{B_c} \log R(B_c)$$

$$\#B_c \log R\,(B_c) = N(B_c)\log R\,(B_c) = \sum_{B_c} \log\frac{y_j}{x_j}$$

$$\log R\,(B_c) = \frac{1}{N(B_c)}\sum_{B_c} \log\frac{y_j}{x_j} = \log\left[\left(\prod_{B_c}\frac{y_j}{x_j}\right)^{1/N(B_c)}\right]$$

$$R(c) = R(B_c) = \left(\prod_{B_c}\frac{y_j}{x_j}\right)^{1/N(B_c)} = \sqrt[N(B_c)]{\frac{y_1}{x_1}\cdots\frac{y_{N(B_c)}}{x_{N(B_c)}}}$$

onde $N(B_c)$ é o recontador (Iglesias, 2019) en B_c, i.e., o seu cardinal. A tabuladoX é a media xeométrica local dos cocientes, de aí a súa denominación. Tamén podémola ver como a razón entre as medias xeométricas locais de cada variábel.

4. EXEMPLOS

No caso da anualización de índices trimestrais pode ser interesante estudarmos o emprego da tabuladoX no canto da media local (Iglesias, 2019), o promediador. Na meirande parte dos casos, a diferenza é irrelevante como p.ex. para toda a economía. Mais pode ter o seu interese cando decemos a un sector pequeno ou a un subsector. No caso do primario temos os seguintes resultados para o valor engadido bruto (VEB)

VEB. Agricultura, gandaría, silvicultura e pesca. Índices de volume encadeados, ano 2010=100

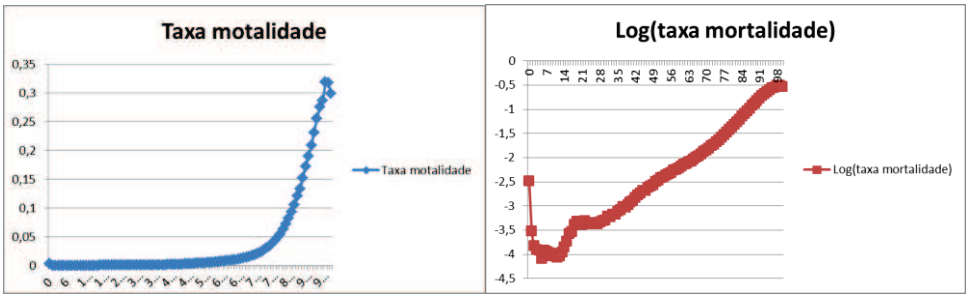
Trimestre	Índice %	Índice	TabuladoX	Promediador	P-X
2010/I	95,38	0,9538			
2010/II	102,15	1,0215			
2010/III	103,11	1,0311			
2010/IV	99,37	0,9937	0,9996	1,0000	0,0005
2011/I	98,05	0,9805			
2011/II	108,80	1,0880			
2011/III	106,38	1,0638			
2011/IV	104,90	1,0490	1,0445	1,0453	0,0008
2012/I	98,06	0,9806			
2012/II	103,14	1,0314			
2012/III	104,54	1,0454			
2012/IV	103,58	1,0358	1,0230	1,0233	0,0003
2013/I	100,51	1,0051			
2013/II	108,95	1,0895			
2013/III	111,86	1,1186			
2013/IV	106,18	1,0618	1,0679	1,0688	0,0008
2014/I	105,00	1,0500			
2014/II	113,95	1,1395			
2014/III	114,34	1,1434			
2014/IV	109,55	1,0955	1,1064	1,1071	0,0007
2015/I	106,73	1,0673			
2015/II	115,61	1,1561			
2015/III	117,48	1,1748			
2015/IV	110,64	1,1064	1,1254	1,1262	0,0008

Fonte: IGE. Contas económicas trimestrais

Se para un estudo máis pormenorizado do sector primario considerarmos o "ano agrario" como o período de 6 trimestres para ter en conta o calendario das diferentes "culturas", i.e, o derradeiro trimestre do ano $a-1$, os 4 trimestres do ano a e o primeiro do ano $a+1$, a diferenza entrambas as dúas aplicacións tabuladoras había acrecentar. Na práctica, en Galiza, sería case que un disparate porque a agricultura pesa pouco e a pesca, moito, no seu sector primario en termos comparativos.

Tamén podémola empregar para alisar as taxas de mortalidade. Dispomos de información da mortalidade por idades simples polo tanto a taxa, $m_{c,a}$ vén referida a un subconxunto do produto cartesiano dos conxuntos $\{0,1,\cdots,c,\cdots,100\}$ e $\{2000,\cdots,a,\cdots 2016\}$. De feito non existen defuncións para algúns anos de determinadas idades. B_c ha ser o subconxunto de anos nos que existen óbitos con c anos cumpridos. $N(B_c)$, en xeral, é igual a 17, mais o seu valor mínimo é 12, i.e., 5 celas "valdeiras". Existe unha ducia de idades con celas valdeiras, $\{3,4,\cdots,14\}$.

Tras a aplicación da tabuladoX, a curva de mortalidade por idades resulta máis lisa. Non existen dous intervalos de monotonía simplemente conexos mais isto pode deberse a motivos sociais (incremento da mortalidade na mocidade nomeadamente nos homes). Temos evitado as mortalidades nulas que aparecen cos datos brutos e alargado os intervalos de monotonía, decrecente $\{0,\cdots,5\}$ e crecente $\{39,\cdots,98\}$; habemos considerar que 100 é a semi-recta derradeira cun comportamento diferente ao resto de idades simples. Graficamente case é imperceptíbel na escala aritmética mais maniféstase nidiamente na escala logarítmica, que é a que se adoita empregar, cuxa primeira diferenza constitúe a imparidade dunha idade á acaroada.



Volvendo ao inicio, podemos empregar a DOL para comparar dúas ratios sen quedármonos só na diferenza de puntos percentuais, p.ex., a fracción de analfabetismo de Galiza e España nos censos de poboación.

Fracción (razón) de analfabetismo. Comparanza sincrónica

Ano censual	Gz	E	Dif. p.%	I	DOL impar.	i Gz/E	i E/Gz
1900	67,85	58,66	9,19	1,1567	0,1455	0,1567	-0,1354
1910	60,05	52,04	8,01	1,1539	0,1432	0,1539	-0,1334
1920	50,28	43,74	6,54	1,1495	0,1393	0,1495	-0,1301
1930	36,24	31,92	4,32	1,1353	0,1269	0,1353	-0,1192
1940	24,41	23,17	1,24	1,0535	0,0521	0,0535	-0,0508
1950	15,86	17,33	-1,47	0,9152	-0,0886	-0,0848	0,0927
1960	11,92	13,51	-1,59	0,8823	-0,1252	-0,1177	0,1334
1970	9,42	8,77	0,65	1,0741	0,0715	0,0741	-0,0690
1981	6,6	6,4	0,2	1,0313	0,0308	0,0312	-0,0303

Fonte: CIEG (actualm. IGE) *A poboación de Galicia. Proxeccións*
Nota: Fracción = (Analfabetos de 10 anos e máis)/(Habitantes de 10 anos e máis)

I = Gz/E Nº índice base España
i Gz/E Incremento relativo de Galiza respecto de España
i E/Gz Incremento relativo de España respecto de Galiza

Observamos unha grande asimetría na TVR (un exemplo de que non é alternada), neste caso as columnas dos incrementos relativos, estes dependen da base empregada (Galiza ou España). Tamén un comportamento diferente entre a variación absoluta (Dif. p.%: diferencia en puntos percentuais) e a relativa (DOL), p.ex., entre o censo de 1940 (1,24; 0,05) e o de 1970 (0,65; 0,07).

CONCLUSIÓNS

A DOL ten un grande valor didáctico porque permite interpretar doadamente resultados da estatística matemática en termos descritivos, numerosos estatísticos como vimos de ver, e mesmo transformacións de variábeis como a escala logit, que resulta a DOL entre a probabilidade dun acontecemento e a do seu complementar.

Este útil permite expresarmos máis axeitadamente resultados máis que coñecidos, p.ex., de tomarmos logaritmos na igualdade $\Pr(B|A)\Pr(A) = \Pr(A|B)\Pr(B)$ obtemos que $\log\Pr(B|A) + \log\Pr(A) = \log\Pr(A|B) + \log\Pr(B)$ e de aí $\log\Pr(B|A) - \log\Pr(A|B) = \log\Pr(B) - \log\Pr(A)$ cuxa interpretación é que a variación relativa (DOL) entre a probabilidade de dous sucesos (non "nulos") é igual á variación entre as probabilidades condicionais recíprocas.

Pode evitar o abuso da TVR cando a base non é canónica ou clara como no exemplo do analfabetismo, transcendendo a variación absoluta en puntos percentuais. Unha vantaxe da DOL é a aditividade que non cumpre a TVR, para ver a súa aplicación no problema de comparanza entre series de diferente frecuencia, mensual, trimestral e anual, consulte López et al. (2020). Ademais se $q = y/x$ entón en dúas situacións a e b, $\mathrm{DOL}(q_a, q_b) = \mathrm{DOL}(y_a, y_b) - \mathrm{DOL}(x_a, x_b)$, que podemos interpretar como que a variación relativa dun cociente é a variación relativa do numerador menos a do denominador.

Tamén permite alisar de xeito elemental cando os datos non estean completamente dispoñibles ou cando o que se precise sexa en termos de erro relativo e non absoluto. P.ex., é necesario cando o que pescudemos é un factor de expansión, i.e., un índice para elevar, polo tanto pode ter interese en contas económicas.

Se non hai razóns da estatística matemática nin das teorías da información ou da decisión, pode que a mellor base sexa 2, lb(·) logaritmo binario, porque duplicar é equivalente a 1 como adoito resulta na TVR, en troques, habíamos perder a posibilidade de comparanza directa con ela.

No entanto, na súa contra convén dicir que non está definida para variables que tomen valores nulos ou negativos e que, na suavización, estea a sobreestimar a variábel ou o indicador obxectivo.

REFERENCIAS

Agresti, A. (2002) *Categorical Data Analysis* 2nd ed. John Wiley & Sons, Hoboken, New Jersey

Aitchison, J. (1982) The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B*, 44, 2, 139-177.

Caamaño Alegre, J. (2008) As desviacións orzamentarias na Xunta de Galicia. *Revista Galega de Economía*, 17, 2, pp. 63-86

Cuadrado (dir.), J.R., T. Mancha y R. Garrido (1998) *Convergencia regional en España. Hechos tendencias y perspectivas*. Colección Economía española, num. 8. Fundación Argentaria-Visor Dis. Madrid

IGE (2022) Táboas de mortalidade completas
https://www.ige.gal/igebdt/selector.jsp?COD=6334&paxina=001&c=0201007&idioma=gl

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

Iglesias Patiño, C.L. (2021) Indicadores sumarios dalgúns tabuladores ou das súas compoñentes. *Actas do XV Congreso Galego de Estatística e Investigación de Operacións*. SGAPEIO. Santiago de Compostela

López, M.E., P. Sánchez y C.L. Iglesias (2020) Monitorización de la coyuntura económica regional através de un indicador sintético. *Revista de estudios regionales*, nº 119, pp.15-41

Prasada Rao (2001) Weighted EKS and generalised CPD methods for aggregation at basic heading level and above basic heading level. OECD-WB. Washington

Uriel, E. e I. Gea (1997) *Econometría aplicada*. Editorial AC.

Vilar, J.M., R. Cao y J.A. Vilar (2009) Informe, I (septiembre). Convenio de colaboración IGE-UdC (documento interno)

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

ANÁLISIS DE EFICIENCIA EN EL SECTOR PÚBLICO USANDO LA OPINIÓN DEL EMPLEADO Y DEL CLIENTE: UNA APLICACIÓN EN EL SISTEMA DE SALUD ESPAÑOL.

Tapia, J.A.[1] y Salvador, B.[2]

[1]jesus.tapia@uva.es, Statistics and Operations Research Departament, University of Valladolid
[2] bonifacio.salvador@uva.es, Statistics and Operations Research Departament, University of Valladolid

RESUMEN

Medir la eficiencia relativa de un conjunto fijo y finito de unidades de servicio-producción, caso de hospitales, es un importante propósito del Análisis Envolvente de Datos (DEA). En este trabajo, ilustramos una manera innovadora de medir esta eficiencia utilizando índices estocásticos de la calidad de estos servicios. Los índices obtenidos a partir de la opinión-satisfacción de los clientes son estimadores, desde el punto de vista estadístico, de la calidad del servicio recibido (outputs); mientras que, la calidad del servicio ofrecido se estima con los índices de opinión-satisfacción de los proveedores de servicio (inputs). La estimación de estos indicadores sólo es posible encuestando a una muestra de clientes y a una muestra de proveedores, en cada servicio. Las eficiencias técnicas obtenidas utilizando los modelos DEA clásicos con los indicadores de calidad estimados, son estimaciones de la eficiencia poblacional, desconocida, que se obtendría si, en cada uno de los servicios fuera posible encuestar a todos sus clientes y a todos sus proveedores. Con el objeto de lograr mejor precisión en la estimación, proponemos resultados para determinar el tamaño de la muestra de clientes y proveedores necesario para que con sus respuestas se pueda lograr una precisión, fijada previamente, en la estimación de la eficiencia poblacional de las unidades de servicio-producción mediante un nuevo intervalo de confianza bootstrap. Usando esta metodología bootstrap e índices de opinión de calidad de dos estudios sociales, uno de médicos y otro de pacientes, analizamos la eficiencia DEA del sistema de salud en España.

Palabras y frases clave: Análisis envolvente de datos, Eficiencia, Sistema de salud, Bootstrap.

REFERENCIAS

Forsund, F. R. (2017) Measuring effectiveness of production in the public sector. Omega, 73: 93-103

Mayston, D.J. (2017) Data envelopment analysis, endogeneity and the quality frontier for public services. Annals of Operations Research, 250(1):185–203.

Tapia, J.A. and Salvador, B. (2022) Data envelopment analysis efficiency in the public sector using provider and customer opinion: An application to the Spanish health system. Health Care Management Science

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

PRIMEIRA APROXIMACIÓN Á IDENTIFICACIÓN E CARACTERIZACIÓN DAS VIVENDAS FAMILIARES EN GALICIA

Del Río Viqueira, M. Isabel[1], Silveira Calviño, Solmary[1] e López Vizcaíno, M. Esther[1]

[1] Instituto Galego de Estatística

RESUMO

Para levar a cabo políticas públicas sobre vivenda é fundamental coñecer o número de vivendas familiares que hai en Galicia, así como, distinguir cales están ocupadas e cales constitúen unha segunda vivenda ou están baleiras.

As fontes de información de orixe principalmente administrativa que achegan datos sobre as vivendas de Galicia, foron construídas con outras finalidades, de aí que teñan moitas fortalezas, pero tamén moitas debilidades que faga necesario adaptalas para que se poidan utilizar con fins estatísticos.

Tendo en conta todas estas fontes que proporcionan información sobre as vivendas, o noso obxectivo é identificar as vivendas familiares en Galicia e asignarlles un identificador único, que prevaleza no tempo, e despois ofrecer características de cada unha delas, como saber quen reside en cada vivenda, a súa superficie, etc.

Palabras e frases chave: Vivenda familiar, distancias entre cadeas, envolventes convexas, record linkage, outliers.

1. INTRODUCIÓN

O coñecemento do número e as características das vivendas familiares en Galicia é fundamental para levar a cabo políticas públicas que teñan como destino os inmobles/vivendas nas que habitan os residendes en Galicia.

A fonte de información máis importante que actualmente proporciona información sobre a poboación, o Padrón Municipal de Habitantes (PMH), non ten uns identificadores únicos para as vivendas. Por outra banda, os ficheiros do Catastro Inmobiliario que proporciona a Dirección General de Catastro do Ministerio de Hacienda, conteñen a Referencia Catastral (RC) como identificador único de vivenda, pero non teñen información sobre se as vivendas están ocupadas, son secundarias ou baleiras. No Catastro tamén se dispón, en principio, da ubicación xeográfica precisa (coordenadas) da vivenda. Ademais, a información catastral debería estar bastante actualizada e libre de erros polas repercusións tributarias.

O principal inconveniente de empregar o Catastro é que a súa información non está enlazada coa poboación residente, non se sabe a priori a relación entre as persoas do Padrón e as vivendas/inmobles de Catastro, é dicir, non se sabe quen vive en cada referencia catastral. Outro problema de Catastro é que nalgúns casos falta a división horizontal que fai que as vivendas dun edificio compartan unha única RC.

O noso obxectivo é identificar as vivendas familiares en Galicia e asignarlles un identificador único, que prevaleza no tempo, e despois ofrecer as características de cada unha delas, como saber quen reside en cada vivenda, a súa superficie, etc.

2. FONTES DE INFORMACIÓN

As principais fontes de información que proporcionan datos sobres as vivendas de Galicia e que se empregarán ao longo deste traballo son as seguintes:

Ficheiros do Catastro inmobiliario

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

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

Ficheiro do PMH

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

Rueiro do Censo Electoral

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

Base de datos sociodemográfica

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

Rexistro dos depósitos de fianza de arrendamentos

O Instituto Galego da Vivenda e o Solo (IGVS) proporciónalle ao IGE un rexistro onde constan as fianzas depositadas polos arrendamentos dos bens inmobles en Galicia, co cal contén información daquelas vivendas que están alugadas, identificadas mediante a súa RC. Nos contratos de arrendamento relativos a vivendas e predios urbanos será obrigatoria a esixencia e prestación de fianza en metálico. Teñen a obriga de depositar esta fianza os arrendadores de vivendas e predios urbanos ante o IGVS.

Como se pode ver, todas as fontes teñen orixe administrativa. Este aspecto é moi positivo, menos custo económico e menos carga de recollida de información sobre a poboación, porén obriga a traballar con fontes construídas para outras finalidades que hai que adaptar para poder chegar aos conceptos estatísticos obxecto de análise. Por exemplo:

- Hai que adaptar os conceptos administrativos aos conceptos estatísticos analizados, neste traballo temos que relacionar o concepto vivenda familiar estatística co concepto inmoble e construción do ficheiro do catastro.
- Solucionar determinadas particularidades dos rexistros usados: no catastro non todos os edificios teñen declarada unha división horizontal en varios inmobles; no PMH cada persoa ten asociada unha única residencia principal, non contempla situacións onde se utilicen dúas vivendas á vez (por exemplo por mobilidade laboral).

A maiores, temos o problema da non sempre empregada harmonización entre as variables dos diferentes rexistros administrativos (interoperabilidade semántica). Neste caso, non existe unha relación entre os códigos de vía do catastro e os do Rueiro do Censo Electoral (empregados no PMH) que nos permita establecer unha relación biunívoca entre as rúas de ambos ficheiros. Teremos que desenvolver esta tarefa para chegar a identificar as vivendas onde reside a poboación de Galicia.

O proceso de traballo que se expón a continuación é o resultado de ir resolvendo situacións como as comentadas.

3. PROCESO PARA A ASIGNACIÓN E PRIMEIRA CARACTERIZACIÓN DE VIVENDAS FAMILIARES

O proceso para a identificación e caracterización das vivendas familiares segue tres pasos, en primeiro lugar definirase un directorio de vivendas a partir dos datos do Catastro Inmobiliario, a continuación farase unha asignación entre as vías do Catastro e as vías do Rueiro do Censo Electoral e por último asignaranse as RC ás persoas do PMH.

Definición dun directorio de vivendas familiares

O Catastro non identifica nin contabiliza vivendas, se non que identifica bens inmobles con uso residencial ou construcións con destino residencial. O ben inmoble é unha unidade xurídica, que o Catastro divide en construcións, en función das peculiaridades destas para definir con precisión as características dos inmobles e poder asignarlle a correspondente valoración catastral. A vivenda, sen embargo, é unha unidade física, tal e como se quere identificar neste traballo. Para chegar a este concepto a partir de vivenda e dende os bens inmobles foi necesario facer determinados proceso previos, en parte baseados no traballo de Enrique et al. (2019).

Para obter o directorio de vivendas a partir da información contida nestes ficheiros partirase dos rexistros de bens inmobles e de construcións. Nos rexistros de bens inmobles especifícase o uso principal ao que se destinan cada un dos bens inmobles: residencial, comercial, industrial, etc. No caso de que nun mesmo ben inmoble se identifiquen usos diversos, Catastro asigna para a totalidade deste o que se considere como uso principal. Por outro lado, no rexistro de construcións identifícanse as construcións que constitúen cada ben inmoble, que á súa vez poden ter diversos destinos, ben iguais ou ben diferentes ao uso principal que se asigna ao ben inmoble. Independentemente do uso principal que se asigna no ficheiro de bens inmobles, neste establécese o destino de cada unha das construcións existentes dentro do ben inmoble. Neste rexistro tamén se establece a tipoloxía construtiva, que informa sobre o tipo de construción de acordo coas súas características arquitectónicas e de finalidades de uso: vivenda colectiva, vivenda unifamiliar, ... Ademais Catastro tamén informa sobre a estrutura da propiedade para cada un dos elementos rexistrados. A estrutura da propiedade pode ser horizontal ou vertical. A propiedade horizontal caracterízase por estenderse de maneira privativa sobre un piso ou local do tipo que sexa. A propiedade vertical é polo contrario aquela que se compón dun único propietario cuxo dereito se estende sobre todos os elementos da finca (Enrique et al., 2019).

Partindo do anterior, defínese a vivenda neste traballo, e no caso de edificios con división horizontal, como o conxunto formado por todas as construcións con destino vivenda que pertence a un mesmo ben inmoble. Considérase vivenda ao ben inmoble con uso residencial ou aquel que conte con polo menos unha construción con destino vivenda. No caso de parcelas cun só ben inmoble residencial con construcións de tipoloxía colectiva, defínese como vivenda ao conxunto de construcións dun ben inmoble que comparten a mesma planta e porta, sendo excluídas aquelas cuxa superficie sexa menor que os 25 m^2.

Partindo da definición anterior, o procedemento para a identificación das vivendas empeza por cruzar a información de bens inmobles coa de construcións para asociar a cada inmoble a construción ou conxunto de construcións físicas pertencentes ao mesmo, e a continuación segue 3 pasos:

1.- Identificación das vivendas unifamiliares:
Estas vivendas teñen as seguintes tipoloxías construtivas en Catastro:

0121: vivendas unifamiliares de carácter urbano, edificación illada ou pareada.
0122: vivendas unifamiliares de carácter urbano, en liña ou mazá cerrada.
0131: edificación residencial rural.

Neste caso identifícase como unha vivenda a agregación de todas as construcións con destino residencial. Esta tipoloxía construtiva está habitualmente rexistrada en réxime de propiedade vertical e con este procedemento, extráese de cada ben inmoble, o espazo dedicado só a vivenda. Desta maneira tamén se consegue identificar aquelas vivendas que se encontran dentro dun ben inmoble con uso principal distinto ao residencial, que sería o caso da Figura 1 onde o ben inmoble ten uso principal agrícola, pero dentro do mesmo hai construcións con uso residencial.

Figura 1: Exemplo de ben inmoble con uso principal agrícola, pero que ten construcións con uso residencial.

Fonte: Figura extraída de Enrique et al. (2019).

2.- Identificación das vivendas colectivas formadas por pisos:

Estas vivendas teñen as seguintes tipoloxías construtivas en Catastro:

0111: vivendas colectivas de carácter urbano, edificación aberta.
0112: vivendas colectivas de carácter urbano, en mazá cerrada.

Trátase dos casos dunha parcela con varios bens inmobles residenciais. Neste caso, cada ben inmoble contabiliza como unha vivenda, tal e como se pode observar na Figura 2 esquerda. Nesta figura o edificio da esquerda componse de catro bens inmobles, tres de uso residencial e un con uso comercial. Neste caso seleccionariamos os tres inmobles con uso residencial.

Figura 2: Exemplo de edificio con varios bens inmobles (esquerda) e de edificio cun único ben inmoble e varias construcións (dereita).

Fonte: Figura extraída de Enrique et al. (2019).

3.- Identificación de vivendas colectivas formadas por pisos sen división horizontal:

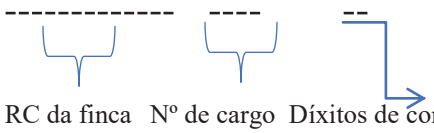
Estas vivendas teñen as seguintes tipoloxías construtivas en Catastro:

0111: vivendas colectivas de carácter urbano, edificación aberta.

0112: vivendas colectivas de carácter urbano, en mazá cerrada.

Este tratamento establécese para as parcelas cun só ben inmoble residencial, con máis dunha construción con destino vivenda e a tipoloxía da vivenda é a de vivenda colectiva. Este sería o caso dun edificio de varias plantas con un só ben inmoble, e que a tipoloxía construtiva de vivenda colectiva indica que pode estar formado por varias vivendas (Figura 2 dereita). Nestes casos, estableceuse que unha vivenda é a agregación das construcións con destino residencial ubicadas no mesmo bloque, escaleira, planta e porta. Contabilizouse unha vivenda por cada dirección diferente a nivel de porta.

O identificador único co que identificaremos cada vivenda será a RC que consta de 20 díxitos coa seguinte estrutura:



En termos xerais, a RC a 14 díxitos identifica un edificio e a RC a 20 díxitos identifica o inmoble/vivenda. No caso das vivendas sen propiedade horizontal, engadirémoslle a RC 4 díxitos máis, que se corresponden co número de construcións dese inmoble.

A dificultade que presenta traballar con estas bases de datos é a dimensión das mesmas, por esta razón neste traballo empregouse o paquete de R dbplyr (Wickham H. et al., 2023).

Procedemento para a unión das vías de Catastro e as vías do Rueiro do Censo Electoral

Establecido o marco de vivendas familiares o seguinte obxectivo é determinar unha relación co Rueiro do Censo Electoral (utilizado no PMH) para poder identificar aquelas vivendas que son residencia habitual da poboación de Galicia. Cada persoa pode ter varias vivendas en propiedade (Catastro) -resida ou non nelas- e ao contrario, unha persoa pode residir (PMH) nunha vivenda que pode ter ou non en propiedade (pode ser alugada ou cedida); polo tanto, a relación entre os ficheiros de Catastro e de PMH deberá efectuarse empregando o enderezo das vivendas.

No PMH as vías do enderezo están codificadas de acordo ao Rueiro do Censo Electoral, porén no caso do Catastro emprégase unha codificación de vías propia, diferente da anterior. O primeiro paso foi

establecer un procedemento para relacionar os códigos das vías do Catastro e os códigos das vías do PMH.

Para cada vía en Catastro o ficheiro inclúe un código, unha variable que indica o tipo de vía (rúa, avenida, praza, etc.) e outra variable que indica o nome da vía. No caso do Rueiro (e no PMH) a identificación das vías en zonas urbanas é similar, hai como en Catastro un código (diferente) e dúas variables que indican o tipo de vía e o seu nome; porén, no caso de zonas rurais estes campos no Rueiro en xeral están en branco (aínda que non sempre), e a localización ven dada polas variables que inclúen o código e o nome da entidade singular e/ou entidade colectiva correspondentes.

Coa información dispoñible, segundo sexa zona urbana ou rural elabórase a "denominación" do enderezo no PMH, que virá dada pola concatenación das variables que definen o tipo de vía e o nome da vía:

- No caso de zona urbana (definida neste procedemento como direccións con código de vía non nulo en PMH), a denominación será o resultado de pegar as variables tipo de vía e nome da vía.
- No caso de zona rural (entendida como aquelas direccións nas que o código da vía toma o valor "00000" ou non ten valor), a denominación virá dada pola unión do tipo de vía e do nome da entidade singular (e/ou entidade colectiva no seu caso).

As denominacións dos enderezos depúranse (suprímense partículas que non aportan información, como espazos en branco, preposicións, símbolos, etc), e posteriormente co paquete de R text2vec (Selivanov D. et al, 2022), que permite comparar textos segundo a distancia do coseno, elaborouse un procedemento que para cada denominación de enderezo incluída no PMH proporciona a denominación de Catastro máis próxima (no sentido de distancia de textos) da forma seguinte:

- Na zona urbana obtense a máis próxima nese mesmo concello, distrito e sección.
- Cos enderezos non resoltos da zona urbana e os aínda non tratados da zona rural obtense a máis próxima no mesmo concello e parroquia.

Determináronse limiares para o valor das distancias que se considerarían como válidas en cada caso.

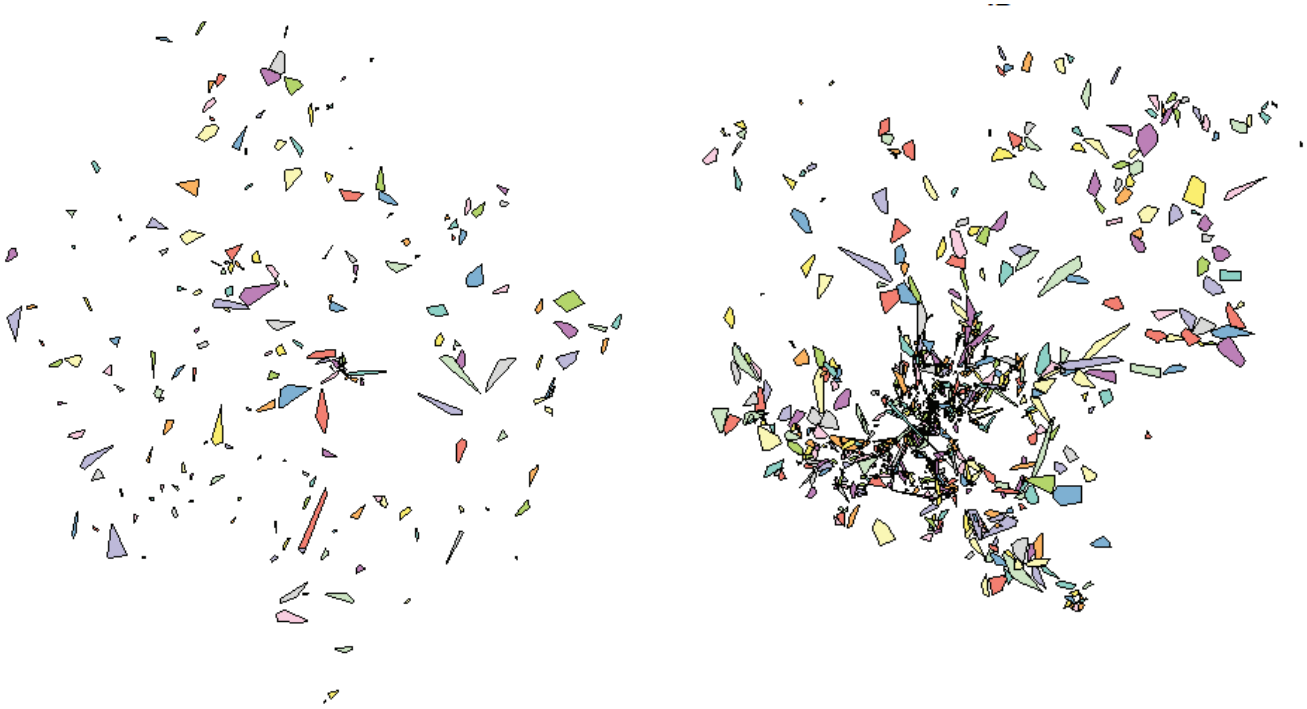
Coa aplicación deste procedemento obtense unha táboa de equivalencias que permite asignar na maior parte dos casos (en torno ao 73% dos casos), a cada vía do Rueiro (e do PMH) unha vía do Catastro.

Para as vías que quedaban sen asignar, e tamén como método de comprobación paras as vías xa asignadas, comprobouse que os puntos das rúas do PMH están incluídas dentro dos polígonos das rúas de Catastro. Para isto seguiuse o seguinte procedemento:

- A partir da Base de datos sociodemográfica determinouse o conxunto de coordenadas xeográficas que pertencían a cada rúa do PMH.
- Para cada concello de Galicia, creáronse os polígonos formados polas envolventes convexas das vivendas de cada rúa de Catastro. Por exemplo, na figura 3 preséntanse os polígonos de rúas do concello de Friol e de Santiago de Compostela.
- Calculouse para cada concello a área dos polígonos.
- Aplicouse o test de Rosner (Rosner, 1983) para eliminar os outliers dos polígonos que tiñan áreas moi grandes e que, polo tanto, cubrían unha boa parte do territorio do concello.
- Comprobouse que a maior parte dos puntos das vías do PMH relacionadas coa vía de Catastro, estaban incluídos nos polígonos de Catastro.
- Se o anterior non ocorre é necesario revisar a correspondencia feita entre a rúa de Catastro e a rúa do PMH.

Figura 3: Polígonos formados polos conxuntos convexos das coordenadas dos edificios dispoñibles no Catastro nos concellos de Friol e de Santiago de Compostela.

FRIOL **SANTIAGO DE COMPOSTELA**



Procedemento para a asignación da RC ao PMH

Neste punto temos dispoñible o directorio de vivendas coas súas características e a súa titularidade, resultado de unir este directorio co rexistro de titularidade de Catastro. Ademais, disponse da relación entre os códigos de vía de Catastro e os códigos de vías do PMH, obtidos no procedemento anterior. Tendo en conta que o 78% das persoas galegas son propietarias da vivenda onde viven (IGE, 2020), cruzaremos o ficheiro da residencia das persoas do PMH co directorio de vivendas de Catastro con titularidade, para asignarlle ás persoas a RC da vivenda onde residen. Neste punto hai que sinalar que nalgunhas persoas do PMH xa tiñamos unha preasignación da RC dun procedemento anterior. Para facer isto seguiremos o seguinte procedemento:

1. Seleccionaremos no directorio de vivendas aqueles propietarios que teñan unha única vivenda por concello.
2. Cruzamos o directorio de vivendas co PMH polo concello e o NIF do propietario.
3. No caso de que as RC dos dous ficheiros coincidan e os códigos de vía do PMH e do Catastro coincidan confirmamos a RC para estes propietarios.
4. No caso de que as RC dos dous ficheiros non coincidan, pero os códigos de vía do PMH e do Catastro coincidan, os números do edificio coincidan (ou sexa baleiro algún deles), cambiamos a RC pola que nos achega o arquivo de Catastro. Se non se compren estas dúas premisas mantemos a RC que tiñamos previamente asignada.

5. No caso de que non teñamos unha asignación previa de RC, os códigos de vía do PMH e do Catastro coincidan e os números dos edificios tamén coincidan (ou sexa baleiro algún deles), asignámoslle a RC de Catastro.
6. No caso no que non haxa unha relación entre os códigos de vía de Catastro e do PMH, facemos unha comparación dos nomes das vías empregando a distancia de Levensthein (Van de Loo, 2014). Cambiamos a RC, se os números coinciden e a distancia de Lenvensthein entre os nomes das rúas é pequena.
7. Seleccionamos agora no directorio de vivendas aqueles propietarios que teñan máis dunha vivenda por concello.
8. Repetimos o procedemento dos puntos 2-6, pero agora tendo coidado de non asignar unha mesma RC a vivendas diferentes.
9. Finalmente, faremos un cruce entre o PMH e o directorio de vivendas empregando os códigos das rúas, a planta e a porta.

Xa por último, empregaremos o ficheiro das fianzas depositadas cando se aluga unha vivenda para completar as RC do ficheiro do PMH. O ficheiro de fianzas permitiranos:

1. Detectar vivendas alugadas que non estaban no directorio de vivendas porque non cumprían os criterios para seren incluídas neste directorio.
2. Asignar a RC ás persoas do PMH que viven en vivendas alugadas e que non a tiñan asignada nos procedementos anteriores. Neste punto haberá que ter coidado, porque tal e como se comentou con anterioridade hai persoas que teñen unha vivenda en propiedade, pero tamén teñen outra/s vivendas alugadas, incluso no mesmo concello.

REFERENCIAS

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

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

Rosner, B. (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. Technometrics. 25, 165-172.

Salcedo, J.A. (2023) Censo de viviendas 2021. Índice. Revista de Estadística y Sociedad, 89, abril 2023.

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

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

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

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

Enquisa sobre vacantes nas empresas galegas, grao de utilización da capacidade produtiva e indicador de sentimento do sector servizos de Galicia. Elaboración e primeiros resultados

Sara Moyano Pérez[1], Ángel Pérez Lago[1]

[1] IGE

RESUMO

A idea de facer estas enquisas xurdiu no Instituto Galego de Estatística ao botar en falta unha serie de vacantes de emprego para a nosa comunidade autónoma coa que poder elaborar unha curva de Beveridge axeitada, que se axuste fielmente á realidade económica de Galicia.

Queriamos tamén reflexar a situación na que se atopaba o emprego no sector servizos, que é o sector maioritario na nosa economía e ter unha idea da medida en que estabamos aproveitando a capacidade produtiva do sector servizos galego.

Estamos polo tanto ante unha operación estatística na que se enmarcan dous estudos:

-Estudo sobre vacantes ou ofertas de emprego nas empresas galegas.

-Estudo da utilización da capacidade produtiva das empresas galegas de servizos e obtención do indicador de sentimento harmonizado.

Quixemos replicar a metodoloxía xa existente na "Encuesta trimestral de costes laborales" (en adiante ETCL) elaborada polo INE e do "Economic Sentiment Indicator" elaborado pola Comisión Europea para poder facer máis adiante comparacións temporais.

Palabras e frases chave: Vacante. Grao de utilización da capacidade produtiva. Indicador de Sentimento Harmonizado.

1. INTRODUCIÓN

Os obxectivos do presente estudo son:

- mellorar a estimación existente do número de postos vacantes nas empresas de Galicia, coa finalidade principal de elaborar unha curva de Beveridge.
- Estimar o grao de utilización da capacidade produtiva do sector servizos.
- Obter un indicador de sentimento harmonizado para o sector servizos.

A curva de Beveridge mostra a relación entre a taxa de vacantes e a taxa de paro. Permite observar a evolución do mercado de traballo. Os desprazamentos ao longo dunha curva teórica son resultado dos efectos do ciclo económico e amosan a conxuntura do mercado: nas fases expansivas a economía móvese cara arriba e cara á esquerda sobre a curva, e realiza o movemento contrario nas fases recesivas. Con todo, se observamos desprazamentos da propia curva teórica estamos ante cambios estruturais no mercado de traballo.

O grao de utilización da capacidade produtiva é un indicador da conxuntura económica que mide indirectamente as variacións da demanda e as presións inflacionarias. É ademais unha das variables necesarias para a estimación da fenda de produción ou "output gap".

A fenda de produción defínese como a diferencia entre o PIB real da economía e o PIB ou produción potencial, e este último defínese á súa vez como o nivel máis alto de produción real que podería ser alcanzado por unha economía sen xerar desequilibrios, se empregase de xeito óptimo os factores produtivos cos que conta. Pode ser negativa, se a economía non alcanza neste momento do ciclo

o seu potencial produtivo, ou positiva, se a demanda induce unha sobreutilización dos factores, xerando presións inflacionistas e problemas de subministro.

Dado o enorme peso do sector servizos no PIB de Galicia e das economías da nosa contorna, tentaremos estimar o grao de utilización da capacidade produtiva deste sector, para o que empregaremos tanto preguntas de medida directa (semellantes ás empregadas no sector industrial) coma preguntas de sentimento económico seguindo a metodoloxía da Comisión Europea.

2. ENQUISA SOBRE VACANTES NAS EMPRESAS GALEGAS

Unha vacante de emprego defínese como segue (Regulamento (CE) No 453/2008 do Parlamento Europeo e do Consello de 23 de abril de 2008 relativo ás estatísticas trimestrais sobre vacantes de emprego na Comunidade):

Vacante de emprego refírese a un posto remunerado creado recentemente ou non ocupado, ou que está a punto de quedar libre, para o que o empregador está tomando medidas activas e está preparado para tomar outras co obxecto de atopar un candidato idóneo alleo á empresa en cuestión é ten a intención de cubrilo inmediatamente ou nun prazo de tempo determinado.

A nivel de comunidade autónoma só tiñamos datos do número de vacantes en xeral e dos motivos polos que non existen postos de traballo vacantes (sen desagregar por sectores, seccións ou tamaño da empresa).

A serie de vacantes que presenta o INE para Galicia é corta e moi irregular; con valores moi baixos respecto aos totais. Se tentamos facer unha curva de Beveridge para Galicia, o resultado é totalmente errático; non temos forma de chegar a ningunha conclusión.

Aproveitando a experiencia das empresas que colaboran na ETCL na nosa comunidade autónoma e calculando unha nova mostra para Galicia esperamos obter uns valores que reflexen máis fielmente a situación real das vacantes así coma un mellor coñecemento dos motivos polos que permanecen ou non eses postos de traballo baleiros; e ter estes datos con celeridade.

Remitimos un cuestionario idéntico ao apartado F da ETCL ("Encuesta trimestral de costes laborales") ás empresas seleccionadas e calculamos os estimadores seguindo a mesma metodoloxía.

Poboación obxecto de estudo: Empresas con polo menos un asalariado no réxime xeral en polo menos un establecemento en Galicia e con actividade incluída nas sección B a S da CNAE.

Variables a investigar:
Número de postos vacantes na empresa ao final do mes de referencia.
Motivo da inexistencia de vacantes no mes de referencia.
Motivo da non cobertura das vacantes do trimestre anterior.

Clasificacións e nomenclaturas: CNAE, Regulamento (CE) No 453/2008 do Parlamento Europeo e do Consello de 23 de abril de 2008 relativo as estatísticas trimestrais sobre vacantes de emprego na Comunidade.

Escalas no tratamento da variables:
O número de vacantes é unha variable numérica que toma valores naturais.
As outras dúas variables son categóricas a escoller entre unha lista que inclúe unha opción aberta.

Unidades estatísticas
Unidades de mostraxe: Empresas incluídas no directorio de empresas e unidades locais do IGE que teñen polo menos un asalariado e con actividade incluída nas sección B a S da CNAE.
Unidade informante: Empresas.
Unidades excluídas: Empresas sen asalariados, así como as correspondentes ás seccións A ,T e U da CNAE.

Método de mostraxe: estratificada con afixación óptima no que a variable de estratificación é o tamaño das unidades definido polo número de traballadores.

Estratos considerados e criterios de estratificación: Consideramos 5 estratos: empresas de 1 a 9 traballadores, de 10 a 49, de 50 a 99, de 100 a 499 e de máis de 500 traballadores. Seleccionaranse todas as empresas do derradeiro estrato (unidades autorrepresentadas).

Determinación do tamaño de mostraxe: calcúlase o tamaño preciso de mostra en cada estrato para a afixación de Neymann.

O resultado, para o ano 2020, cunha precisión global do 5% e un mínimo de 21 empresas por estrato é o seguinte:

Estrato	De 1 a 9 traballadores	De 10 a 49 traballadores	De 50 a 99 traballadores	De 100 a 499 traballadores	Máis de 500 traballadores
nh	167	88	21	48	54
Nh	64769	7137	726	405	54

Tamaño total da mostra: 378 empresas.
A mostra repartirase en tres submostras mensuais.
Cada ano renóvase un 20% da mostra, exceptuando as unidades do último estrato que se investigan de modo exhaustivo.

Marco: Directorio de empresas do IGE.

Procedemento de estimación
Estimaremos o total de vacantes mediante o estimador separado de razón, empregando como variable auxiliar o número de traballadores no directorio de empresas do IGE.

O estimador separado de razón é igual á suma dos estimadores de razón en cada un dos estratos.

$$\hat{X}_R = \sum_h \hat{X}_{Rh}$$

Onde:

$$\hat{X}_{Rh} = \hat{R}_h D_h = \frac{\sum_{i=1}^{n_h} X_{hi}}{\sum_{i=1}^{n_h} D_{hi}} D_h = \frac{\sum_{i=1}^{n_h} X_{hi}}{d_h} D_h$$

Sendo:

h: estrato definido polo tamaño das empresas galegas (número de traballadores).
Nh: tamaño poboacional para o estrato h.
nh: tamaño mostral para o estrato h.
Xhi: valor da variale X na empresa i do estrato h.
Dhi :número de traballadores no marco da empresa i do estrato h.

O total de traballadores no marco das empresas do estrato h virá dado por:

$$D_h = \sum_{i=1}^{N_h} D_{hi}$$

E o total de traballadores no marco das empresas do estrato h pertencentes á mostra será:

$$d_h = \sum_{i=1}^{n_h} D_{hi}$$

Agrupando os términos que acompañan a cada valor Xhi observado nun estrato h temos que o factor de elevación dentro de cada estrato h é:

$$F_h \equiv \frac{D_h}{d_h}$$

Estimaremos tamén as frecuencias das distintas causas de inexistencia de vacantes e de non cobertura das existentes co mesmo procedemento de estimación.

3. ESTUDO DA UTILIZACIÓN DA CAPACIDADE PRODUTIVA DAS EMPRESAS GALEGAS DE SERVIZOS E INDICADOR DE SENTIMENTO HARMONIZADO DO SECTOR SERVIZOS

Para estimar as taxas de crecemento potenciais e a fenda de produción a Comisión Europea utiliza indicadores directos de utilización da capacidade produtiva para o sector industrial, pero para os servizos e a construción emprega indicadores de sentimento económico no seu lugar, principalmente por non dispoñer de datos directos.

O INE pregunta pola utilización da capacidade produtiva do sector industrial na elaboración dos indicadores de confianza empresarial, pero non contamos con datos directos para o resto da economía.

Dado o enorme peso do sector servizos no PIB de Galicia e das economías da nosa contorna, tentaremos estimar o grao de utilización da capacidade produtiva deste sector, para o que empregaremos tanto preguntas de medida directa (semellantes ás empregadas no sector industrial) coma preguntas de sentimento económico seguindo a metodoloxía da Comisión Europea.

Preguntaremos, por unha banda, pola utilización dos factores produtivos da empresa (traballadores e instalacións, maquinaria e equipamento) no trimestre en curso, así como polas expectativas para o trimestre seguinte. E preguntaremos tamén pola impresión da marcha da empresa durante o trimestre e pola percepción da evolución da demanda, realizando as preguntas que a Comisión Europea emprega para os indicadores de sentimento económico.

Poboación obxecto de estudo: Empresas con polo menos un asalariado no réxime xeral en polo menos un establecemento en Galicia e con actividade incluída nas seccións G a S da CNAE, exceptuando as seccións O, P e Q.

Ámbito temporal: A información recollerase trimestralmente e pedirase información relativa ao trimestre en curso e expectativas para o seguinte trimestre.

Variables a investigar:
Grao de ocupación do persoal no trimestre en curso
Previsión de ocupación do persoal para o próximo trimestre.
Grao de utilización das instalacións, maquinaria e equipamento no trimestre en curso.
Previsión de utilización das instalacións, maquinaria e equipamento para o próximo trimestre.
Percepción da evolución da situación da empresa nos últimos tres meses.
Percepción da evolución da demanda de servizos da empresa nos últimos tres meses.
Previsión da evolución da demanda de servizos da empresa nos próximos tres meses

Escalas no tratamento de variables:
Aínda que o grao de ocupación do persoal e das instalacións, maquinaria e equipamento se cuantifican en porcentaxes, os informantes so teñen que escoller o intevalo en que sitúan estes, polo que todas as variables que medimos son categóricas.

Unidades estatísticas
Unidades informantes: Empresas seleccionadas na mostra.
Unidades excluídas: Co fin de obter valores comparables a escala europea excluiremos do marco as empresas sen asalariados e as empresas con actividade comprendida nas seccións O (administracións públicas), P (educación) e Q (sanidade e servizos sociais) da CNAE.

Deseño mostral
Método de mostraxe: estratificada con afixación óptima no que a variable de estratificación é o tamaño das unidades definido polo número de traballadores.

Consideramos 5 estratos: empresas de 1 a 9 traballadores, de 10 a 49, de 50 a 99, de 100 a 499 e de máis de 500 traballadores. Seleccionaranse todas as empresas do derradeiro estrato (unidades autorrepresentadas).

Determinación do tamaño de mostraxe: calcúlase o tamaño preciso de mostra en cada estrato para a afixación de Neymann. Impoñemos un mínimo de 20 elementos en cada estrato para evitar que se eleve o erro de mostraxe.

O resultado para o ano 2020, cunha precisión global do 5% é o seguinte:

Estrato	De 1 a 9 traballadores	De 10 a 49 traballadores	De 50 a 99 traballadores	De 100 a 499 traballadores	Máis de 500 traballadores
nh	270	103	20	43	25

Tamaño total da mostra: 461 empresas

Marco: Directorio de empresas do IGE.

Procedemento de estimación

Intensidade de uso da capacidade produtiva: dado o carácter exploratorio das preguntas, estudaremos a distribución de frecuencias de cada unha das repostas, deixando para o futuro a elaboración de indicadores baseados nos valores observados.

Para o cálculo do indicador de sentimento do sector servizos basearémonos na Enquisa de servizos harmonizada UE e replicaremos o cálculo dos índices empregados pola Comisión Europea.

Para o cálculo do indicador de utilización de capacidade combinado, a Comisión Europea utiliza tres elementos: unha medida directa de utilización da capacidade industrial e dous indicadores de sentimento económico, un para a construción e outro para o sector servizos. Replicaremos o cálculo do indicador de sentimento económico para o sector servizos (economic sentiment indicator for the services sector ou ESI.SERV).

O cálculo realízase coma segue:

Primeiramente calcularemos a porcentaxe de respostas (positivas, neutras e negativas) para cada estrato. A porcentaxe farémola dun xeito simple (número de respostas de cada tipo sobre o número total de empresas no estrato).

O resultado da variable X no estrato h e para un mes t é un vector columna :

Xh =(Ph , Eh , Mh)

Sendo:

Ph porcentaxe de empresas con incremento

Eh porcentaxe de empresas que non cambian

Mh porcentaxe de empresas que diminúen

Obteremos así o resultado para cada estrato e un resultado final para o total das empresas galegas que será unha media ponderada dos resultados en cada estrato. O coeficiente de ponderación reflictirá a significación ou peso relativo de cada estrato dentro da poboación total de empresas. Así:

$X= (\sum Ph\ Wh\ , \sum Eh\ Wh\ , \sum Mh\ Wh\ )$ siendo $\sum Wh =1$

Wh é o coeficiente de peso relativo para cada estrato

$\sum Ph\ Wh + \sum Eh\ Wh +, \sum Mh\ Wh = 100$

O balance (B), o saldo final, calcularémolo:

B = P - M

Recollida e tratamento de datos de ambos estudos

Os datos recolleranse mediante un cuestionario web a cubrir pola empresa.

O contacto inicial coa empresa será por correo postal, mediante unha carta dirixida ao enderezo que consta no directorio de empresas na que se incluirá a chave de acceso para cubrir o cuestionario.

A recollida, gravación, depuración e validación da información levarase a cabo mensualmente. Nunha primeira fase remítese o cuestionario por correo ás unidades informantes xunto coa carta de presentación. A continuación, realízanse roldas de contactos telefónicos coas empresas nas que non se obtivo a resposta ou nas que se considera insuficiente ou dubidosa a recibida.

No proceso de recollida da información pódense presentar incidencias que impidan ou dificulten a obtención dos cuestionarios cubertos: negativas, pechadas, non localizadas. Para poder atender a determinadas incidencias sobre a mostra inicial, sen que se minore o tamaño de deseño da mostra, efectúanse substitucións. A selección de substitutas efectúase dentro do estrato correspondente á empresa substituída.

O procedemento de depuración e validación dos datos lévase a cabo nas seguintes fases: depuración manual, depuración informática e validación posterior da calidade.

Os criterios de validación e depuración reducen a falta de resposta, tanto global ó conxunto do cuestionario como parcial a apartados específicos. Nos casos nos que non fora posible obter información procederase á súa imputación a partir doutros cuestionarios de características parellas ou de fontes complementarias de información.

A publicación dos resultados ten unha periodicidade trimestral.

A presentación dos resultados para o estudo das vacantes ou ofertas de emprego nas empresas galegas e para medir a utilización da capacidade produtiva no sector servizos, farémola en formato de táboa multidimensional, contemplando unha agregación por tamaño da empresa.

O indicador de sentimento presentará un saldo ou balance xeral para o sector servizos e un desglose segundo o tamaño das empresas de servizos .

4. PRINCIPAIS RESULTADOS

Actualmente temos publicados datos dende o primeiro trimestre de 2022 e neste 2023 estamos facendo resumos trimestrais.

O segundo trimestre de 2023 deixounos os seguintes resultados:

No segundo trimestre de 2023, o 23,3% das vacantes requirían titulación STEM. Esta porcentaxe sube con respecto ao mesmo trimestre do ano anterior, no que o 13,9% das vacantes requirían titulación STEM.

As empresas con maior número de asalariados son as que teñen máis vacantes e as máis pequenas as que teñen menos necesidade de cubrir postos vacantes.

O principal motivo polo que as empresas non tiñan vacantes foi “Non se necesita ningunha persoa máis”, que superou o 79% no segundo trimestre de 2023.

No segundo trimestre de 2023 o 30,3% das empresas con vacantes aínda tiñan sen cubrir as do trimestre anterior. Esta porcentaxe foi do 20,9% no segundo trimestre do ano anterior.

O motivo máis indicado en todos os trimestres como causante de non ter cubertas as vacantes foi a falta de persoas con formación axeitada (sen ter en conta a formación STEM e a capacitación dixital); a porcentaxe foi de 29,5% no segundo trimestre, algo inferior a do mesmo trimestre do ano anterior (41,3%).

No segundo trimestre de 2023, o 56,4% das empresas empregan as súas instalacións, maquinaria e/ou equipamento máis do 75%. Esta porcentaxe diminuíu con respecto do mesmo trimestre do ano anterior, no que a porcentaxe de empresas que empregaba as súas instalacións, maquinaria e/ou equipamento máis do 75% era do 60,7%.

No que respecta ao grao de ocupación do persoal, no segundo trimestre de 2023, o 67,5% das empresas empregan ao seu persoal máis do 75%.

A percepción nas empresas galegas de servizos da situación empresarial mantense en valores positivos nos últimos catro trimestres. O mesmo ocorre coa percepción da evolución da demanda de servizos.

A previsión da evolución da demanda para o próximo trimestre acada o valor 20,02 no segundo trimestre de 2023, e mantén un valor positivo coma no trimestre anterior.

REFERENCIAS

European Commission (2021) On The Joint Harmonised EU Programme of Business and Consumer Surveys.

Havik Kare, Mc Morrow Kieran, Orlandi Fabrice, Planas Christophe, Raciborski Rafal,Werner Röger, Rossi Alessandro, Thum-Thysen Anna, Vandermeulen Valerie (2014) On Economic papers 535. The production function methodology for calculating potencial growth rates & output gaps.

ESTUDIO LONGITUDINAL Y ANÁLISIS DE SUPERVIVENCIA APLICADO AL MERCADO LABORAL DE GALICIA

Noa Veiguela Fernández[1], María Esther López Vizcaíno[2]

[1] Instituto Galego de Estatística (IGE)
[2] Instituto Galego de Estatística (IGE)

RESUMEN

La Muestra Continua de Vidas Laborales (MCVL) constituye una fuente de información fundamental para estudiar el mercado laboral de Galicia desde una óptica longitudinal y retrospectiva. En el año 2022 el Instituto Galego de Estatística (IGE) publicó una estadística sobre la vida laboral de la población afiliada a la Seguridad Social en Galicia, que ha permitido analizar aspectos como:

- la edad de incorporación de los gallegos y las gallegas al mercado laboral
- la duración de sus períodos de afiliación
- a qué edad se experimenta el primer episodio de desempleo
- el porcentaje de la vida laboral que se ha estado cotizando a la Seguridad Social;

y, lo más importante, ha permitido comparar cómo han variado estos indicadores desde la década de los 60 hasta la actualidad. Los datos retrospectivos de la MCVL también han permitido aplicar las técnicas del análisis de supervivencia, muy usadas en demografía y epidemiología, al estudio de las transiciones del empleo al desempleo y viceversa. Mediante estas técnicas hemos obtenido estimaciones del riesgo al que se enfrentan las personas que trabajan en Galicia de perder su empleo y hemos analizado si este riesgo ha ido en aumento con el paso de los años o, por el contrario, se ha reducido. En esta ponencia se presentan los principales resultados obtenidos de ambos análisis, el longitudinal y el de supervivencia.

Palabras y frases clave: MCVL, análisis longitudinal, análisis de supervivencia, estimador de Kaplan-Meier, función de supervivencia, función de riesgo

1. INTRODUCCIÓN

La Muestra Continua de Vidas Laborales (en adelante, MCVL) es un conjunto de microdatos anonimizados procedentes de diversos registros administrativos: de las bases de datos de la Seguridad Social, del Padrón municipal de habitantes del Instituto Nacional de Estadística (INE) y del resumen anual de retenciones e ingresos a cuenta del IRPF (Modelo 190) de la Agencia Tributaria. Constituye una muestra representativa[1] de todas las personas que han tenido relación con la Seguridad Social durante el año de referencia en España, bien porque han estado afiliadas en situación de alta laboral, bien porque han percibido alguna prestación o subsidio por desempleo, o bien porque han percibido algún tipo de pensión contributiva de la Seguridad Social (por ejemplo, la de jubilación) (Seguridad Social, 2020). La MCVL sigue el devenir laboral de las personas que la integran, incorporando en cada nueva edición información del mismo grupo poblacional (el 4% de la población afiliada en 2004, momento de extracción de la primera edición); esta peculiaridad la convierte en una fuente ideal para realizar análisis longitudinales del mercado laboral español. Puede ser considerada también un conjunto de información de carácter retrospectivo, ya que no sólo incluye información del grupo observado desde 2004 en adelante: incorpora

[1] En IGE (2019) se ha contrastado la representatividad de la MCVL para el caso gallego.

también información anterior a este año que figure en las bases de datos de la Seguridad Social para cada uno de sus integrantes.

Estas particularidades de la MCVL son las que la distinguen de otras fuentes estadísticas, en las que se dispone de información para distintos momentos del tiempo, pero procedente de unidades de observación distintas (fuente de tipo transversal) o bien, aun procediendo del mismo grupo, los períodos de observación no son continuos (fuente de tipo longitudinal pero recogida a intervalos discretos de tiempo). La MCVL, como su nombre indica, ofrece información continua durante todo el período en que uno de sus integrantes tiene relación con la Seguridad Social, quedando constancia de los sucesivos vínculos contractuales que mantiene a lo largo de su vida laboral, así como de los períodos de desempleo; también registra información relativa a interrupciones de los episodios de empleo por baja laboral, maternidad, etc., así como el fin de la vida profesional por pase a situación de jubilación. La línea cronológica sólo se interrumpe cuándo la persona deja de tener relación con este organismo.

Para poder ofrecer información longitudinal, hay que someter a depuración los ficheros originales que distribuye la Seguridad Social (ya que contienen incoherencias entre episodios de empleo y desempleo; Veiguela et al, 2014) y transformarlos en una base de datos tipo panel. Este procedimiento, en el que no nos vamos a detener en esta ponencia, se ha realizado con el apoyo del programa de gestión de grandes conjuntos de datos, SQL Server Management, y con el software estadístico R[2]. El resultado es una tabla como la que se presenta a continuación, donde se ha condensado la información que figura en los ficheros originales de la MCVL para el sujeto de ejemplo E.

Las variables *estado*, *episodio_tipo* y *episodio*, que no vienen en los ficheros originales, se han añadido a la base de datos longitudinal para facilitar la comprensión de la información:

- *estado* indica el tipo de episodio de que se trata; sus valores son *empleo*, *desempleo* y *no_relacion*, esta última categoría para indicar los períodos de tiempo en que "se pierde la pista" del sujeto en los ficheros de la Seguridad Social.
- *episodio_tipo* indica el orden del episodio dentro de todos los de su tipo, esto es, si se trata del primer episodio de empleo, del segundo, del tercero, etc., del primer episodio de desempleo, del segundo, etc.
- *episodio* indica el orden del episodio en el cómputo total, sin distinguir de qué tipo se trate (empleo o desempleo), sin tener en cuenta los episodios de falta de relación.

Tabla 1: Ejemplo de cómo se ha estructurado la base de datos longitudinal creada a partir de la MCVL. Edición 2018

id_per	fecha_alta	fecha_baja	num_dias	estado	episodio_tipo	episodio	sexo
E	07/10/1976	29/10/1976	23	empleo	1	1	Mujer
E	30/10/1976	01/11/1976	3	no_relacion	1	0	Mujer
E	02/11/1976	10/11/1976	9	empleo	2	2	Mujer
E	11/11/1976	28/11/1976	18	no_relacion	2	0	Mujer
E	29/11/1976	19/12/1976	21	empleo	3	3	Mujer
E	20/12/1976	27/12/1976	8	no_relacion	3	0	Mujer
…							
E	11/07/2016	22/12/2016	165	empleo	33	38	Mujer
E	23/12/2016	05/01/2017	14	no_relacion	34	0	Mujer
E	06/01/2017	25/06/2017	171	desempleo	6	39	Mujer
E	26/06/2017	20/12/2017	178	empleo	34	40	Mujer
E	21/12/2017	31/12/2017	11	no_relacion	35	0	Mujer
E	01/01/2018	31/12/2018	365	desempleo	7	41	Mujer

continuación columnas…

[2] Puede consultarse el procedimiento empleado por el IGE para depurar y transformar los ficheros originales en una base de datos adecuada para su explotación longitudinal en IGE (2022), bajo petición expresa, empleando el siguiente formulario web:
https://www.ige.gal/web/peticioninfo.jsp?idioma=gl

continuación columnas…

fecha_nacimiento	edad	cnes	…	cnae09	num_relaciones
01/09/1962	56	31	…	00	1
01/09/1962	56	31	…	-	0
01/09/1962	56	31	…	00	1
01/09/1962	56	31	…	-	0
01/09/1962	56	31	…	00	1
01/09/1962	56	31	…	-	0
…					
01/09/1962	56	31	…	78	1
01/09/1962	56	31	…	-	0
01/09/1962	56	31	…	-	0
01/09/1962	56	31	…	78	8
01/09/1962	56	31	…	-	0
01/09/1962	56	31	…	-	0

Fuente: elaboración propia a partir de los ficheros *AFILIAD* y *PERSONAL* correspondientes a la edición 2018 de la MCVL.

Nota: (-) No procede

Por ejemplo, en la vida laboral del sujeto E (tabla 1), del que sabemos que es una mujer nacida el 01/09/1962, se sucede un primer episodio de empleo, que dura 23 días (fila 1: *estado=empleo, episodio_tipo=1 y episodio=1*); a continuación, el sujeto desaparece de los ficheros suministrados por la Seguridad Social durante 3 días (fila 2: *estado=no_relacion, episodio_tipo=1 y episodio=0*), trascurridos los cuales es contratado de nuevo, dando inicio su segundo vínculo laboral (fila 3: *estado=empleo, episodio_tipo=2 y episodio=2*). De esta forma, podemos seguir su trayectoria desde 1976 hasta el 31/12/2018, que es el momento en el que finaliza el período de estudio (ya que, para la construcción de la base de datos longitudinal, se han utilizado los microdatos de la MCVL correspondientes a la edición 2018). En la trayectoria profesional de esta persona, que se prolongó durante 42 años, se sucedieron hasta 41 episodios distintos y no siempre continuados, de empleo y desempleo. De hecho, el último año de observación, la persona lo pasó íntegramente en situación de desempleo, percibiendo una prestación o subsidio por este concepto; esta era la séptima vez que la persona del ejemplo se encontraba en esta situación (última fila: *estado=desempleo, episodio_tipo=7 y episodio=41*).

Finalmente, para simplificar el análisis de los resultados, se ha decidido considerar como un mismo episodio de empleo todos los vínculos laborales de la persona que se sucedan de forma ininterrumpida en el tiempo. En la base de datos longitudinal se incorpora la información del vínculo que se considere más representativo del periodo, en base al siguiente criterio:

- el episodio de mayor duración
- en caso de empate en antigüedad, el episodio más cercano al momento de finalización del estudio

Para identificar cuántos vínculos laborales distintos integran cada una de las filas de empleo de la tabla, se ha incluido la variable *num_relacions*. Por ejemplo, en el caso del sujeto E, se observa que su última relación laboral (que corresponde al trigésimo-cuarto episodio de empleo de la persona y que tiene lugar entre el 26/06/2017 y el 20/12/2017) resume 8 vínculos laborales distintos, que tienen lugar de forma ininterrumpida (*num_relacions=8*).

A partir de la información contenida en esta base de datos longitudinal, se ha elaborado la estadística *Muestra continua de vidas laborales. Información de carácter longitudinal (cohortes de edad)*, cuyas aportaciones más interesantes al estudio del mercado laboral gallego se presentan en el apartado siguiente. Los resultados completos pueden consultarse en la página web del IGE, en la siguiente dirección: https://www.ige.gal/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0204034002

2. EXTRACCIÓN DE INDICADORES DESCRIPTIVOS DEL MERCADO LABORAL A PARTIR DE LA BASE DE DATOS LONGITUDINAL

Para estudiar cómo ha ido evolucionando el mercado laboral gallego con el paso del tiempo, se ha agrupado a los integrantes de la MCVL longitudinal en 3 conjuntos, en función de su año de nacimiento: los y las que han nacido antes de 1969 (tienen 50 o más edad en 2018; integran la cohorte 1), las personas que han nacido entre 1969 y 1983 (cumplieron entre 35 y 49 años en 2018; cohorte 2), y la cohorte 3, integrada por las personas de menor edad, con menos de 35 años (han nacido en 1984 o con posterioridad a esta fecha).

Como se observa en la tabla 2, el grupo con mayor presencia en la base de datos es el de 35 a 49 años, que representa el 42,64% de la muestra longitudinal; el 32,73% de las personas cuenta con 50 o más años y el grupo con menor representación es el de menos de 35 años, el 24,63% de la muestra total. En esta tabla se presentan los principales indicadores sobre la inserción laboral de la población de Galicia en función de la cohorte de nacimiento.

Tabla 2: Resumen de los principales indicadores sobre la inserción laboral de la población de Galicia en función de su cohorte de nacimiento

Indicador	Total	Cohorte 1: >=50 años	Cohorte 2: 35-49 años	Cohorte 3: <=34 años
Peso en la muestra (%)	100,00	32,73	42,64	24,63
Edad media de ingreso al mercado laboral (años)	22,9	24,9	22,5	20,8
Duración media del primer episodio de empleo (años)	2,8	5,0	2,1	0,9
Duración media de los periodos de empleo (años)	4,1	7,2	3,4	1,2
Número medio de episodios de empleo (episodios)	14,4	14,1	16,4	11,3
Densidad media de cotización posible (%)	52,3	59,6	56,5	35,1
Densidad media de cotización real (%)	72,5	76,6	75,2	62,2
Tiempo medio transcurrido hasta experimentar el primer episodio de desempleo (años)	6,3	8,0	5,8	4,5
Edad media a la que se experimenta el primer episodio de desempleo (años)	28,1	31,1	27,4	24,3
Duración media del primer episodio de desempleo (años)	0,5	0,6	0,5	0,4

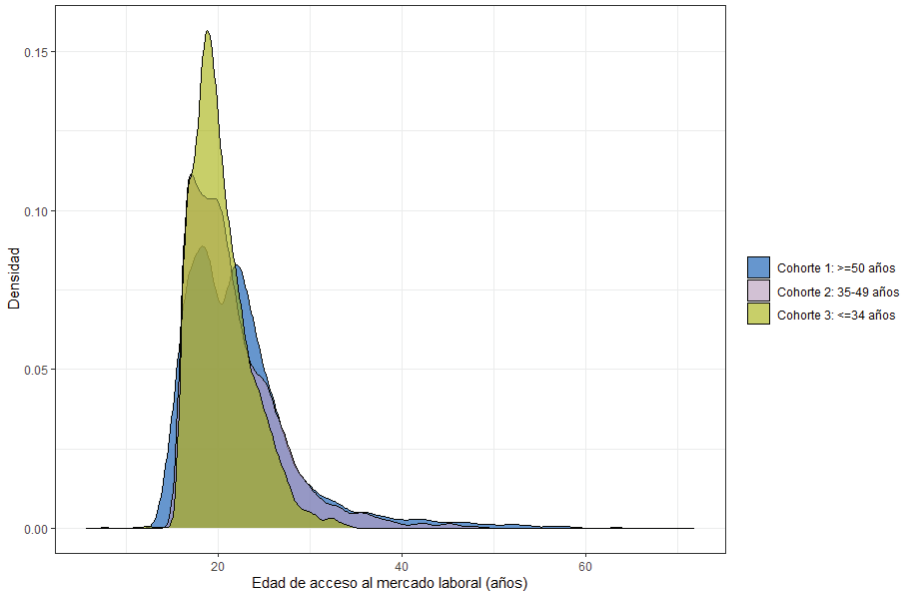
Fuente: IGE. Muestra continua de vidas laborales. Información de carácter longitudinal (cohortes de edad)

Nota: para el cálculo del indicador *densidad media de cotización posible* se divide el tiempo total en situación de alta laboral de la persona por la vida laboral posible de la cohorte, si esta hubiese accedido al mercado de trabajo en el momento de cumplir los 16 años (edad mínima legal para trabajar actualmente en España). Por el contrario, en el cálculo del indicador *densidad media de cotización real* se tiene en cuenta el momento de acceso al mercado laboral de cada persona; es decir, se divide su vida laboral por el tiempo transcurrido desde que se produce el ingreso hasta el 31/12/2018.

Como se observa en la tabla 2, la edad media de ingreso al mercado laboral ha disminuido desde 1960 en Galicia: las personas de menor edad comienzan a trabajar antes de lo que lo habían hecho sus padres y madres y sus abuelos y abuelas. En las figuras 1 y 2 se representan las distribuciones de la variable edad de acceso al primer empleo, por cohortes de nacimiento, para hombres y mujeres de forma separada; se observa que, en todos los casos, esta distribución está sesgada hacia la derecha: la mayor parte de la población comienza a trabajar entre los 19 y los 25 años. No obstante, la distribución de los nacidos y de las nacidas con posterioridad a 1983 (cohorte 3) está más concentrada que la correspondiente a la cohorte 2 y esta, a su vez, más que la cohorte 1. Así, la edad media a la que comienza a trabajar la

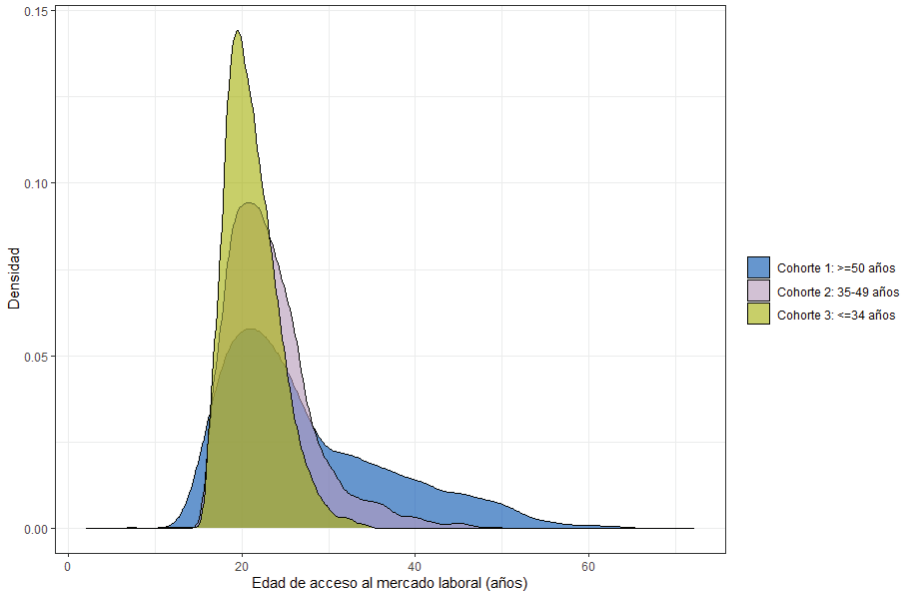
población con menos de 35 años en 2018 se sitúa en 20,8 años, sube a 22,5 para las personas que cumplen 35 a 49 años y alcanza los 24,9 años para la cohorte de mayor edad.

Figura 1. Distribución de la edad a la que se accede al primer empleo según la cohorte de nacimiento. Hombres



Fuente: IGE. Muestra continua de vidas laborales. Información de carácter longitudinal (cohortes de edad)

Figura 2. Distribución de la edad a la que se accede al primer empleo según la cohorte de nacimiento. Mujeres



Fuente: IGE. Muestra continua de vidas laborales. Información de carácter longitudinal (cohortes de edad)

Sospechamos que estas diferencias por grupos pueden estar relacionadas en cierta manera con la progresiva incorporación de la mujer al mercado laboral, ya que la distribución de la edad de ingreso de

las mujeres de mayor edad (cohorte 1 – figura 2) está mucho menos concentrada que la relativa a los varones y que las correspondientes a las cohortes femeninas 2 y 3. Este hecho confirmaría que la mujer "mayor" se ha incorporado al trabajo (o mejor dicho, ha comenzado a cotizar al Sistema de la Seguridad Social) de forma muy paulatina y a edades avanzadas.

La duración media de los episodios de empleo es mayor, como cabe esperar, en el grupo de población de mayor edad: en esta cohorte sus episodios laborales duran, de media, unos 7,2 años, frente a los 3,4 años para la cohorte 2 y los 1,2 años para la 3. Es lógico pensar que los primeros contratos sean los más inestables y, por ello, los de más corta duración, mientras los trabajadores y las trabajadoras aún se están formando y buscando mejorar sus condiciones laborales; a medida que avanza la carrera profesional, se tiende a la estabilidad laboral y los episodios de empleo van creciendo en duración. No obstante, si nos fijamos, no en la duración media de todos los episodios, sino únicamente en la del primer empleo, llama la atención cómo esta ha ido disminuyendo en los registros de la Seguridad Social: cuando las personas nacidas antes de 1969 entraron a formar parte del mercado laboral gallego, su primer contrato tenía una duración media de 5 años; para los nacidos y las nacidas entre 1969 y 1983, esta duración había bajado ya a los 2,1 años; y para las personas que cumplieron menos de 35 años en 2018 (esto es, las nacidas con posterioridad a 1983), recién llegadas al mundo laboral, su primer episodio duró menos de 1 año (unos 11 meses). Estas diferencias por cohortes en el valor de esta variable parecen indicar que está aumentando la rotación laboral en Galicia, al menos en los primeros años de la carrera profesional.

El indicador *densidad media de cotización posible* se calcula poniendo en relación el número de años que cada persona de la muestra lleva afiliada a la Seguridad Social en relación al cómputo total posible de su generación (considerando como fecha de inicio de esta los 16 años, que es la edad legal actual a la que se puede comenzar a trabajar en España). Por su parte, la *densidad media de cotización real* relaciona el número de años en activo con lo que le correspondería a cada persona si se mantuviese en esta situación a lo largo de toda su vida laboral. La diferencia entre un indicador y otro estriba en que en el primer caso el cálculo de la vida laboral total parte de los 16 años, mientras que en el segundo comienza cuando la persona da inicio a su vida laboral (que puede coincidir o no con los 16 años, aunque, en media, se sitúa en torno a los 23 años – tabla 2).

En media, la población de Galicia ha pasado trabajando el 52,3% de la vida laboral completa que le hubiese correspondido de haber accedido al mercado laboral a los 16 años y el 72,5% de su vida laboral real (teniendo en cuenta la verdadera edad a la que cada persona comenzó a trabajar). Por cohortes, los indicadores aumentan con la edad del grupo considerado: la cohorte 1 ha trabajado el 59,6% de su vida laboral posible y el 76,6% de la real, la cohorte 2 el 56,5% (75,2% en el caso de la vida laboral real) y, para la cohorte 3 (los y las más jóvenes), el indicador se sitúa muy lejos del 50%, en el 35,1% (supera ligeramente el 60% en el caso del segundo indicador).

La población de Galicia tarda unos 6,3 años de media en experimentar su primer episodio de desempleo, una vez que comienza a trabajar. Los valores de esta variable han disminuido con el paso de los años, de tal forma que hoy en día la falta de empleo se experimenta antes que en décadas previas: una persona que hubiese comenzado su andadura profesional en los años 60 tardaba 8 años de media en convertirse en parada por primera vez (cohorte 1); en los setenta y primera mitad de los 80, este valor había bajado a 5,8 años (cohorte 2). La población que ingresa al mercado laboral a partir de 1985 experimenta su primera situación de desempleo 4,5 años después de haber comenzado a trabajar.

En relación con lo que acabamos de mencionar, la edad media a la que los gallegos y las gallegas experimentamos nuestro primer episodio de desempleo está disminuyendo, como cabía esperar: de media, "caemos en paro" por vez primera a los 28,1 años, pero este valor es inferior en la cohorte de menor edad (que se quedan sin empleo en torno a los 24) y mayor que la media, por el contrario, para la cohorte de mayor edad (31,1 años para los de 50 o más). Por el contrario, la duración de los períodos de desempleo es mayor en este último grupo poblacional: estas personas pasaron unos 7 meses de media en paro (0,6 años) después de su primer episodio laboral. Para el grupo de 35 a 49 años el valor baja a 6 meses (0,5 años) y los de menor edad son los pasan menos tiempo en paro entre trabajo y trabajo (en torno a 4 meses). De estos resultados se extrae que, en la actualidad, la población gallega experimenta más joven el paso a la situación de desempleo de lo que lo hacían sus mayores, pero se tarda menos tiempo en abandonar este estado y encontrar un nuevo empleo.

3. APLICACIÓN DE LAS TÉCNICAS DEL ANÁLISIS DE SUPERVIVENCIA A LA INFORMACIÓN DE LA BASE DE DATOS LONGITUDINAL DE LA MCVL

Las técnicas del *análisis de supervivencia* permiten estudiar la variable "tiempo hasta que ocurre un suceso" y su dependencia de posibles variables explicativas[3]. Aplicado al mercado laboral, este tipo de análisis podría emplearse para obtener indicadores como el tiempo que transcurre hasta que una persona pasa del empleo al desempleo, el riesgo que tiene de quedarse en paro, etc. Existen infinidad de técnicas alternativas que permiten abordar el estudio de estas variables, como el *análisis de la varianza* o los *modelos de regresión*, pero presentan dos inconvenientes frente al análisis de supervivencia (HU Ramón y Cajal, 2007):

- La variable "tiempo hasta que ocurre un evento" no sigue, por regla general, una distribución normal, hipótesis de partida en este conjunto de técnicas.
- El estudio de esta variable exige la observación del fenómeno durante un periodo de tiempo prolongado, en el que suelen ocurrir *pérdidas*, esto es, momentos del tiempo en los que no se tiene información de los sujetos observados; por ejemplo, en el caso que nos ocupa, el mercado laboral estudiado a partir de la relación de los sujetos con la Seguridad Social, una *pérdida* correspondería al momento de tiempo para el que no se tiene información de la persona en las bases de datos de este organismo.

La variable "tiempo hasta que ocurre un evento" es una variable aleatoria continua, no negativa, cuya *función de probabilidad o de densidad* viene dada por (HU Ramón y Cajal, 2007; Sociedad Colombiana Cardiología y Cirugía Cardiovascular, 2017):

$$P(t_1 \leq T \leq t_2) = \int_{t_1}^{t_2} f(t)dt \qquad 0 < t < \infty$$

La expresión anterior indica la probabilidad de que el tiempo de espera hasta que suceda el evento observado se sitúe entre el momento t_1 y t_2. En este contexto, la *función de distribución o de probabilidad acumulada* mide la probabilidad acumulada hasta un momento concreto del tiempo:

$$P(t_1 \leq T) = \int_{0}^{t_1} f(t)dt \qquad 0 < t < \infty$$

La *función de supervivencia* se define como la probabilidad complementaria de la anterior:

$$S(t_1) = 1 - P(t_1 \leq T) = \int_{t_1}^{\infty} f(t)dt \qquad 0 < t < \infty$$

y expresaría la probabilidad de que el tiempo de espera hasta que ocurre el evento observado se sitúe después del momento t_1. Otra función interesante en el contexto del análisis de supervivencia es la *función de riesgo*, que es la función de densidad de T condicionada a que $T \geq t_1$; es decir, indica qué probabilidad hay de que, si el evento no ha ocurrido hasta el momento t_1, esto es, el tiempo de espera ha llegado hasta el momento t_1, el evento suceda en ese momento concreto de tiempo:

$$h(t_1) = P(T = t_1 / T \geq t_1) = \frac{P(T = t_1 \cap T \geq t_1)}{P(T \geq t_1)} = \frac{f(t_1)}{S(t_1)}$$

En el contexto del mercado laboral, si, por ejemplo, estamos analizando el tiempo que una persona está trabajando y comparándolo con el momento en que deja de trabajar y pasa a la situación de desempleo, el evento observado (el evento *rupturista*), que supone la interrupción del periodo de observación, sería "caer en paro"; el tiempo de espera hasta que ocurre el evento observado sería el

[3] Este conjunto de técnicas se ha empleado mucho en demografía y en sanidad; por ejemplo, en el estudio de las enfermedades crónicas, el tiempo hasta que ocurre la muerte del enfermo o de la enferma (*tiempo de supervivencia*) y su dependencia de la aplicación de distintos tratamientos (Castro, 2018; HU Ramón y Cajal, 2007; Boj, 2017; USC-FEGAS). También se han empleado en el control de la calidad o durabilidad de un producto para determinar, por ejemplo, el tiempo hasta que este producto se estropea (*tiempo de fallo*). No obstante, son ya varios los investigadores que han detectado la potencialidad de estas técnicas y las han comenzado a aplicar en sus estudios sobre el mercado laboral español, como en Alonso (2018) y Bentolilla et al. (2018).

tiempo que la persona pasa trabajando (tiempo en el que la persona "sobrevive", tiempo de supervivencia al evento observado).

Cuando se trabaja con información procedente de una muestra de personas, en particular cuando se analizan variables relacionadas con sus trayectorias laborales, que exigen la observación durante un periodo dilatado, hay que tener en cuenta la posibilidad de que se produzcan *pérdidas*; es decir, se deje de tener información de la persona durante el periodo en estudio. Existen dos motivos por los cuales pueden aparecer estas pérdidas (HU Ramón y Cajal, 2007):

- Debido al fin del estudio, esto es, que el periodo de observación acabe sin que se haya producido el evento "rupturista"; por ejemplo, si fijamos el periodo de observación en 2 años, todas aquellas personas con un contrato que exceda este número serán consideradas pérdidas al final del estudio.
- Debido a la pérdida de la persona, de la que dejamos de tener información durante el periodo de observación. Cuando una relación laboral finaliza, si la persona tiene derecho a la percepción de una prestación o subsidio por desempleo, lo lógico es que después del episodio de empleo aparezca un episodio de desempleo en los ficheros de la Seguridad Social. O bien, si no tiene derecho a este tipo de percepción o la finalización de un contrato se debe a que la persona ha encontrado otro empleo que la satisfaga más, después del primer episodio laboral se producirá un segundo episodio, sin que medie demasiado tiempo entre uno y otro. Pero, en ocasiones, la persona desparece de los ficheros de la Seguridad Social durante un periodo de tiempo muy largo, durante el cual no se sabe si la persona está trabajando o no (puede cotizar por medio de un sistema distinto de provisión al de la Seguridad Social o incluso trabajar "en negro") o si sigue formando parte de la población objetivo (ha podido cambiar de domicilio y dejar, con ello, de formar parte de esta; puede que haya cumplido la edad de jubilación y abandonado, por tanto, el mercado laboral; puede incluso que se haya producido el fallecimiento de la persona).

Para estimar el tiempo de espera hasta la ocurrencia de un suceso, así como sus funciones de densidad, de distribución, de supervivencia y de riesgo y, siempre que se cuente con una muestra de tamaño considerable, se puede recurrir al *método de Kaplan-Meier* (HU Ramón y Cajal, 2007; Sociedad Colombiana Cardiología y Cirugía Cardiovascular, 2017). Se trata de un procedimiento de estimación no paramétrico: no se asume ninguna función de probabilidad para la variable en estudio, sino que se estima ésta a partir de las observaciones de la muestra por el método de máxima verosimilitud, empleando la tabla de frecuencias de la variable en estudio (a aquello que es más frecuente, es decir, a aquello que es más verosímil que ocurra en base a las observaciones disponibles, le corresponderá una mayor probabilidad teórica de ocurrir). Este método, además, permite incluir la información perdida en el estudio, partiendo de la hipótesis de que, si se produce una pérdida, el evento ocurre en un momento posterior al momento en el que se observa la pérdida (Boj, 2017; Castro, 2018).

Supongamos que contamos con una muestra aleatoria simple de tamaño considerable n (la MCVL cumple con esta premisa) extraída de la población objetivo, en la cual se observa el tiempo hasta la ocurrencia de un determinado suceso T, cuyos valores vendrán dados por la secuencia $t_1 < t_2 < t_3 < \cdots < t_k$, donde $k \leq n$. Para cada tiempo t_i existen n_i individuos en riesgo de experimentar el suceso en ese momento (elementos de la muestra en los cuales el evento aún no ha ocurrido, por lo que para estas personas $T \geq t_i$); en ese momento t_i sabemos, además, que se observan d_i eventos, es decir, d_i personas que experimentan el suceso en ese momento concreto del tiempo. Además, se asume que durante el intervalo $[t_i, t_{i+1})$ se producen m_i pérdidas.

La función de verosimilitud de la muestra viene dada por:

$$L = \prod_{i=1}^{k} h_i^{d_i}(1 - h_i)^{n_i - d_i}$$

donde h_i expresa la función de riesgo en el momento t_i ($h(t_i)$). Maximizando la función anterior, se obtiene el estimador de máxima verosimilitud de la función de riesgo:

$$\widehat{h_i} = \frac{d_i}{n_i} \qquad i = 1,2,3,\dots,k \qquad \text{(fórmula de cálculo 1)}$$

La estimación de la función de supervivencia viene dada, a su vez, por…

$$\widehat{S(t_i)} = \prod_{j|t_j<t_i}^{i} \left(1 - \frac{d_j}{n_j}\right) \qquad i = 1,2,3,\ldots,k$$

(fórmula de cálculo 1)

Además, si la muestra de observaciones no presenta censuras, en ese caso los *estimadores de Kaplan-Meier* se pueden obtener directamente a través de las estimaciones de las funciones de densidad y distribución, calculadas a partir de la tabla de frecuencias de la variable (HU Ramón y Cajal, 2007; Sociedad Colombiana Cardiología y Cirugía Cardiovascular, 2017). La estimación de la función de densidad vendría dada por…

$$\widehat{f_i} = \frac{d_i + m_i}{n} \qquad i = 1,2,3,\ldots,k$$

La estimación de la función de distribución se calcularía como el sumatorio de las frecuencias relativas o, lo que es lo mismo, de las estimaciones de la función de densidad:

$$\widehat{F(t_i)} = \sum_{j|t_j<t_i}^{i} \widehat{f_i} \qquad i = 1,2,3,\ldots,k$$

Por lo tanto, siempre y cuando la tabla de frecuencias no presente censuras u observaciones perdidas, la estimación de la función de supervivencia también se puede obtener como el complementario de la estimación de la función de distribución…

$$\widehat{S(t_i)} = 1 - \widehat{F(t_i)}$$

(fórmula de cálculo 2)

y la función de riesgo, a su vez, como…

$$\widehat{h_i} = \frac{\widehat{f_i}}{\widehat{S(t_i)}}$$

(fórmula de cálculo 2)

Ayudándonos de las librerías *survival* (Therneau, 2023) y *survminer* (Kassambara et al., 2021) del software libre R, aplicamos el análisis de supervivencia al estudio de la permanencia en el primer episodio laboral de cada una de las personas que integran la muestra longitudinal de la MCVL (Martínez, 2017; Lastra, 2019; Urdinez et al., 2021). Hemos aplicado estas técnicas de forma separada para cada uno de los 3 grupos en que se dividió la muestra en el apartado anterior, teniendo en cuenta la cohorte de nacimiento, para observar si existen diferencias en los resultados estimados en función del momento en que se accediese al mercado laboral. En la tabla 3 se presentan las *estimaciones de Kaplan-Meier* de las funciones de supervivencia y riesgo en el primer empleo para cada cohorte, en la que…

t_i = tiempo (en días) de permanencia en el primer empleo
n_i = personas en riesgo de perder su empleo al finalizar el día considerado
d_i = eventos observados, esto es, personas que han perdido su empleo al finalizar el día considerado
m_i = censuras, pérdidas de información (personas que desaparecen de la muestra en el día considerado)
f_i = estimación de la función de densidad; es decir, estimación de la probabilidad de que un individuo pierda su empleo al finalizar el día considerado
S_i = *estimación de Kaplan-Meier* de la función de supervivencia; es decir, estimación de la probabilidad de que un individuo permanezca en su empleo, al menos, al finalizar el día considerado
h_i = *estimación de Kaplan-Meier* de la función de riesgo; es decir, estimación de la probabilidad de que un individuo conserve su empleo hasta el día antes del considerado y justo lo pierda al finalizar este

La probabilidad de que una persona de la muestra pierda su primer empleo al día siguiente de comenzar a trabajar es de 0,0369, mientras que la probabilidad de que lo pierda en el día 31 (después de trabajar un mes entero) es de 0,0157 (columna *fi*). Visto desde otra óptica, la probabilidad de que una persona "sobreviva" a su primer día de empleo es de 0,9631 y de que "sobreviva" al primer mes de trabajo es de 0,7950 (columna *Si*). Expresado en porcentajes, en torno al 3,7% de las personas pierden su primer empleo al día siguiente de comenzar en él y el 1,6% sólo ocupan su primer puesto de trabajo durante un mes. Por su parte, el 96% de las personas de la muestra sobrevive a su primer día de trabajo y el 79% al primer mes. La probabilidad de que una persona que ha trabajado durante todo un año (fila 365 días de la tabla 3) pierda su primer empleo al día siguiente es de 0,0672 (columna *hi*). Esta probabilidad

es elevada si se compara con la de los días previos y posteriores, lo que tiene relación con el hecho de que muchos contratos se firman con una duración máxima prevista de 1 año.

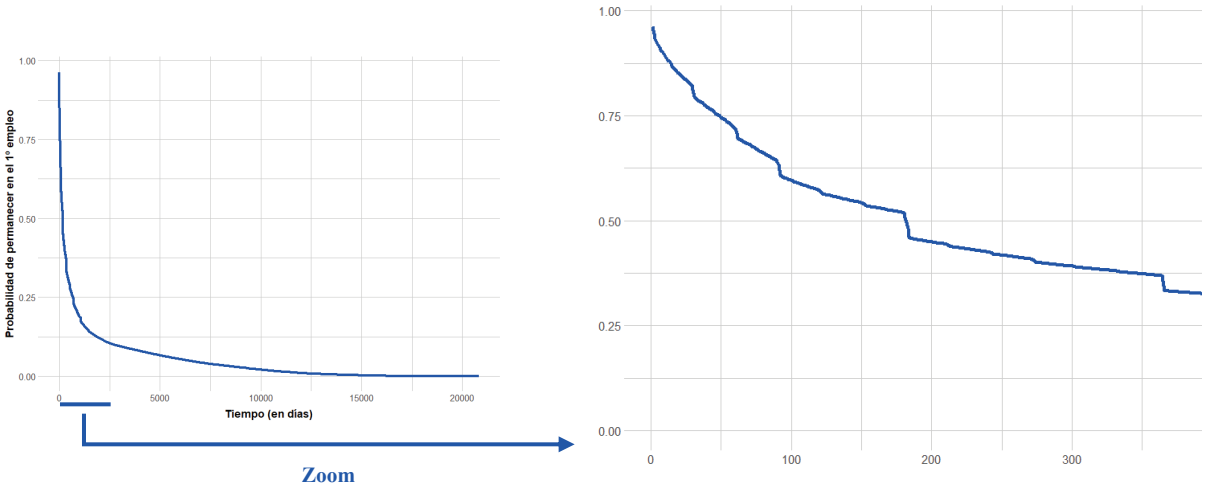
En las figuras 3 y 4 se resumen las funciones de supervivencia y de riesgo estimadas y presentadas en la tabla 3. Se ha hecho *zoom* sobre el intervalo de tiempo que va desde el día 1 al 400 para resaltar la estrecha relación que existe entre ambas funciones. La primera, la de supervivencia, sigue una trayectoria descendente desde el primer día, pero esta no es lineal; se produce una caída pronunciada de la supervivencia en el primer empleo en torno al día 182 (6 meses) y, de nuevo, en torno al día 365 (cuando se cumple un año de contrato). Esto se debe a que muchas relaciones laborales dan comienzo con una fecha de finalización determinada de antemano y, normalmente, las fechas límite suelen coincidir con meses enteros (contratos de un mes, seis meses o un año). A su vez, la función estimada de riesgo sigue una trayectoria ascendente no lineal, con picos que coinciden con los momentos en los que la probabilidad de supervivencia registra bruscas caídas, en torno al día 182 o 365; en estos momentos, el riesgo de perder el empleo es elevado debido a la propia estructura del mercado laboral gallego.

Tabla 3: Tabla de frecuencias de los tiempos (en número de días) de permanencia en el primer episodio laboral de los integrantes de la muestra longitudinal de la MCVL. Edición 2018, con la *estimación de Kaplan-Meier* de las funciones de supervivencia y riesgo

ti	ni	di	mi	fi	Si	hi
1	47.525	1.753	-	0,0369	0,9631	0,0369
2	45.772	853	-	0,0179	0,9452	0,0186
3	44.919	530	-	0,0112	0,9340	0,0118
4	44.389	375	-	0,0079	0,9261	0,0084
5	44.014	392	-	0,0082	0,9179	0,0089
6	43.622	252	-	0,0053	0,9126	0,0058
7	43.370	251	-	0,0053	0,9073	0,0058
8	43.119	248	-	0,0052	0,9021	0,0058
9	42.871	251	-	0,0053	0,8968	0,0059
10	42.620	229	-	0,0048	0,8920	0,0054
…	…	…	…	…	…	…
28	39.431	158	-	0,0033	0,8264	0,0040
29	39.273	215	-	0,0045	0,8218	0,0055
30	39.058	532	-	0,0112	0,8106	0,0136
31	38.526	744	-	0,0157	0,7950	0,0193
32	37.782	153	-	0,0032	0,7918	0,0040
…	…	…	…	…	…	…
179	24.729	29	-	0,0006	0,5197	0,0012
180	24.700	60	-	0,0013	0,5185	0,0024
181	24.640	600	-	0,0126	0,5058	0,0244
182	24.040	513	-	0,0108	0,4950	0,0213
183	23.527	648	-	0,0136	0,4814	0,0275
184	22.879	971	-	0,0204	0,4610	0,0424
185	21.908	116	-	0,0024	0,4585	0,0053
…	…	…	…	…	…	…
363	17.545	23	-	0,0005	0,3687	0,0013
364	17.522	28	-	0,0006	0,3681	0,0016
365	17.494	1.176	-	0,0247	0,3434	0,0672
366	16.318	476	-	0,0100	0,3333	0,0292
367	15.842	48	-	0,0010	0,3323	0,0030
…	…	…	…	…	…	…
20668	2	1	-	0,0000	0,0000	0,5000
20819	1	1	-	0,0000	-	1,0000

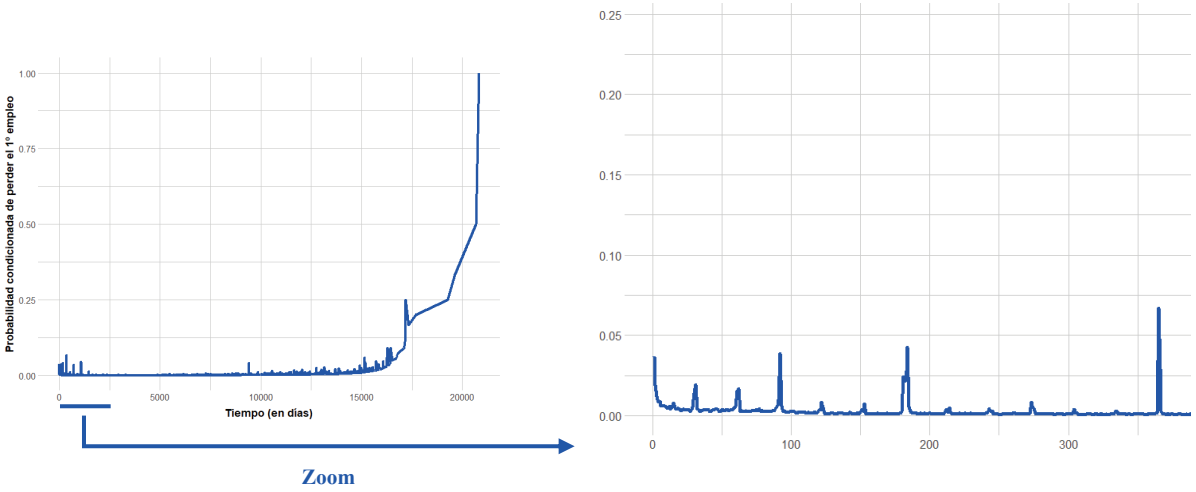
Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

Figura 3. *Estimación de Kaplan-Meier* de la función de supervivencia en el primer episodio laboral de los integrantes de la muestra longitudinal de la MCVL



Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

Figura 4. *Estimación de Kaplan-Meier* de la función de riesgo de perder el primer empleo para los integrantes de la muestra longitudinal de la MCVL



Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

La comparación de las funciones estimadas de supervivencia y riesgo de cada grupo de edad en su primer vínculo laboral (tabla 4) nos permite investigar si se está precarizando la entrada en el mercado laboral gallego. La probabilidad de permanencia en el primer empleo (función de supervivencia) es mayor en la cohorte 1 (la de mayor edad en 2018), que en la cohorte 2 y la de este grupo, a su vez, superior que en la cohorte 3, cualquiera que sea el intervalo de tiempo considerado. El 99,26% de las personas que hoy cuentan con 50 o más edad "sobrevivieron" a su primer día de trabajo; entre la población de 35 a 49 años, este porcentaje se situó en 96,34% y para el grupo de menor edad, los que cumplieron menos de 35 años en 2018, bajó al 92,29%. Tras el primer mes de contrato, el 90,27% de la cohorte 1 seguía trabajando; en la cohorte 2 este porcentaje había bajado al 80,62% y en la cohorte 3 no llegaba al 70% (69,45%). Es decir, en la década de los 60 y comienzos de los 70, 9 de cada 10 personas que ingresaban al mercado laboral mantenía su primer contrato durante más de 30 días; a finales de los 70 y primera mitad de los 80, 8 de cada 10 personas superaban los primeros 30 días de relación laboral; a partir de la segunda mitad de los 80 y ya en los 90, esta relación había bajado a 7 de cada 10.

Las diferencias entre cohortes se acentúan al considerar la supervivencia al primer año de contrato: el 48,90% del grupo 1 seguía contratado después de haber cumplido el primer año de trabajo; para la cohorte 2, este porcentaje se sitúa más de 10 puntos por debajo (el 31,64% de las personas que hoy tienen entre 35 y 49 años estuvo contratada más de un año en su primer empleo). Pero el mayor desfase se observa en la cohorte 3, la de menor edad, con una probabilidad de "supervivencia" en el primer empleo de sólo 0,1945; este valor implica que sólo el 19,45% de los jóvenes permanece en su primer trabajo durante al menos un año.

Tabla 4: Tabla de frecuencias de los tiempos (en número de días) de permanencia en el primer episodio laboral de los integrantes de la muestra longitudinal de la MCVL. Edición 2018, con la *estimación de Kaplan-Meier* de las funciones de supervivencia y riesgo, por cohortes de edad

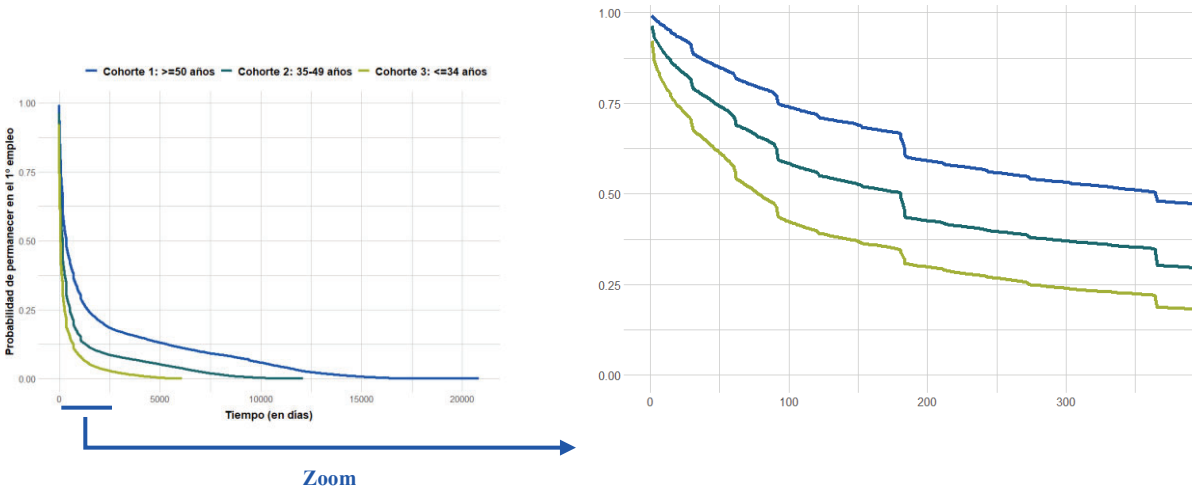
ti	Cohorte 1: >=50 años			Cohorte 2: 35-49 años			Cohorte 3: <=34 años		
	fi	Si	hi	fi	Si	hi	fi	Si	hi
1	0,0074	0,9926	0,0074	0,0366	0,9634	0,0366	0,0771	0,9229	0,0771
2	0,0044	0,9881	0,0045	0,0187	0,9447	0,0195	0,0348	0,8882	0,0377
3	0,0038	0,9843	0,0039	0,0113	0,9334	0,0120	0,0207	0,8674	0,0233
4	0,0036	0,9807	0,0036	0,0082	0,9251	0,0088	0,0131	0,8543	0,0151
5	0,0040	0,9767	0,0041	0,0084	0,9167	0,0091	0,0136	0,8407	0,0160
6	0,0033	0,9733	0,0034	0,0060	0,9107	0,0065	0,0067	0,8340	0,0080
7	0,0033	0,9700	0,0034	0,0048	0,9060	0,0052	0,0088	0,8251	0,0106
8	0,0026	0,9674	0,0027	0,0055	0,9005	0,0061	0,0082	0,8170	0,0099
9	0,0019	0,9655	0,0020	0,0065	0,8939	0,0073	0,0076	0,8094	0,0093
10	0,0029	0,9626	0,0030	0,0053	0,8886	0,0059	0,0066	0,8028	0,0081
…	…	…	…	…	…	…	…	…	…
28	0,0024	0,9174	0,0026	0,0034	0,8217	0,0042	0,0043	0,7120	0,0060
29	0,0035	0,9139	0,0038	0,0044	0,8173	0,0053	0,0061	0,7059	0,0086
30	0,0112	0,9027	0,0123	0,0111	0,8062	0,0135	0,0114	0,6945	0,0161
31	0,0151	0,8875	0,0168	0,0152	0,7910	0,0189	0,0171	0,6774	0,0246
32	0,0032	0,8843	0,0036	0,0030	0,7880	0,0038	0,0036	0,6738	0,0054
…	…	…	…	…	…	…	…	…	…
179	0,0004	0,6682	0,0006	0,0005	0,5039	0,0010	0,0011	0,3475	0,0032
180	0,0019	0,6663	0,0028	0,0010	0,5029	0,0020	0,0009	0,3466	0,0027
181	0,0119	0,6545	0,0178	0,0154	0,4875	0,0306	0,0088	0,3378	0,0254
182	0,0115	0,6430	0,0175	0,0127	0,4748	0,0261	0,0065	0,3313	0,0192
183	0,0135	0,6295	0,0210	0,0161	0,4587	0,0339	0,0095	0,3218	0,0287
184	0,0222	0,6073	0,0352	0,0224	0,4363	0,0488	0,0147	0,3072	0,0456
185	0,0047	0,6026	0,0078	0,0017	0,4346	0,0038	0,0007	0,3065	0,0022
…	…	…	…	…	…	…	…	…	…
363	0,0008	0,5049	0,0015	0,0005	0,3488	0,0014	0,0001	0,2202	0,0004
364	0,0004	0,5045	0,0008	0,0008	0,3480	0,0023	0,0005	0,2196	0,0024
365	0,0156	0,4890	0,0309	0,0316	0,3164	0,0907	0,0251	0,1945	0,1143
366	0,0076	0,4814	0,0155	0,0132	0,3032	0,0418	0,0077	0,1868	0,0395
367	0,0020	0,4794	0,0041	0,0007	0,3025	0,0023	0,0003	0,1866	0,0014
…	…	…	…	…	…	…	…	…	…

Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

Al comparar las funciones estimadas de riesgo, es la cohorte 3 la que presenta un mayor riesgo de pérdida del primer empleo en todo el intervalo de tiempo analizado. La probabilidad de que el contrato finalice al año de dar comienzo se sitúa en 0,1143 para la población menor de 35 años, en 0,0907 para el grupo de 35 a 49 años y alcanza el valor más bajo entre las personas de 50 o más edad (0,0309). Por lo tanto, parece que, efectivamente, la inserción de las sucesivas generaciones de gallegos y gallegas en el mercado laboral se está "precarizando" con el paso del tiempo, entendiendo este adjetivo como una situación en la que la probabilidad de firmar contratos de larga duración está disminuyendo. En

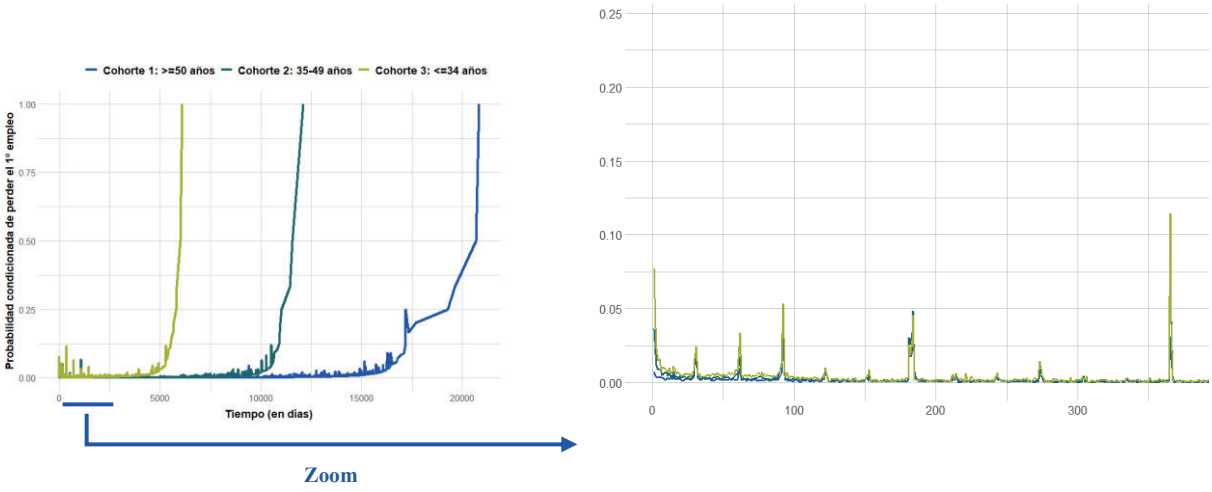
contraposición, el riesgo de que el primer empleo finalice en un corto intervalo de tiempo no ha dejado de aumentar desde la década de los 70 (figuras 5 y 6).

Figura 5. *Estimación de Kaplan-Meier* de la función de supervivencia en el primer episodio laboral de los integrantes de la muestra longitudinal de la MCVL, por cohortes de edad



Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

Figura 6. *Estimación de Kaplan-Meier* de la función de riesgo de perder el primer empleo para los integrantes de la muestra longitudinal de la MCVL, por cohortes de edad



Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

Por último, decidimos aplicar también el análisis de supervivencia al estudio del tiempo que una persona pasa en situación de desempleo, luego de experimentar su primer vínculo laboral (es decir, se ha modelizado la duración del primer episodio de desempleo). Se fijó como tiempo máximo de estudio 365 días, que es el momento a partir del cual el desempleo adquiere la connotación de paro de larga duración. Por lo tanto, un episodio de estas características, que se prolongue durante más de un año, habrá que considerarlo una observación perdida por fin de estudio[4]. En la tabla 5 y figura 7 se resumen los resultados de este análisis, particularizado también por cohortes de edad.

[4] El hecho de que existan observaciones censuradas implica que las fórmulas de cálculo 1 y 2 de los *estimadores de Kaplan-Meier* para las funciones de supervivencia y riesgo no ofrecen los mismos

De nuevo, la cohorte que representa a la población de menor edad presenta las estimaciones más bajas de la función de supervivencia en el primer episodio de desempleo, cualquiera que sea el intervalo de tiempo considerado. Pero en este caso, que los valores de esta sean inferiores debe considerarse una buena señal, ya que implican una menor probabilidad de permanencia en situación de paro durante periodos prolongados de tiempo. Así, en la cohorte 3, el 87,60% de las personas estuvo en esta situación durante más de un mes, lo que quiere decir que el 12,40% de la cohorte ya había abandonado el colectivo de desempleados en este tiempo. En la cohorte 2, la probabilidad de permanecer en situación de desempleo durante, como mínimo, 30 días, se sitúa en 0,9073 y, para la cohorte 1, este valor sube a 0,9210; en este último grupo, sólo el 7,9% de sus integrantes encontró un empleo durante los 30 primeros días en paro.

Si nos fijamos en las últimas filas de la tabla 5, las diferencias entre cohortes se acentúan, pero en este caso, son favorables para el grupo de menor edad: sólo en el 5,28% de las observaciones pertenecientes al grupo de los menores de 35 años el primer episodio de desempleo se convirtió en paro de larga duración. En la cohorte 2 (aquellos y aquellas que ingresaron al mercado laboral en la década de los 70 y primera mitad de los 80) el 12,78% de las observaciones se convirtieron en paro de larga duración y en la cohorte 1 (ingresaron al mercado laboral antes de 1970), en 1 de cada 4 casos (el 25,17%) el primer episodio de paro acabó convirtiéndose en desempleo de larga duración.

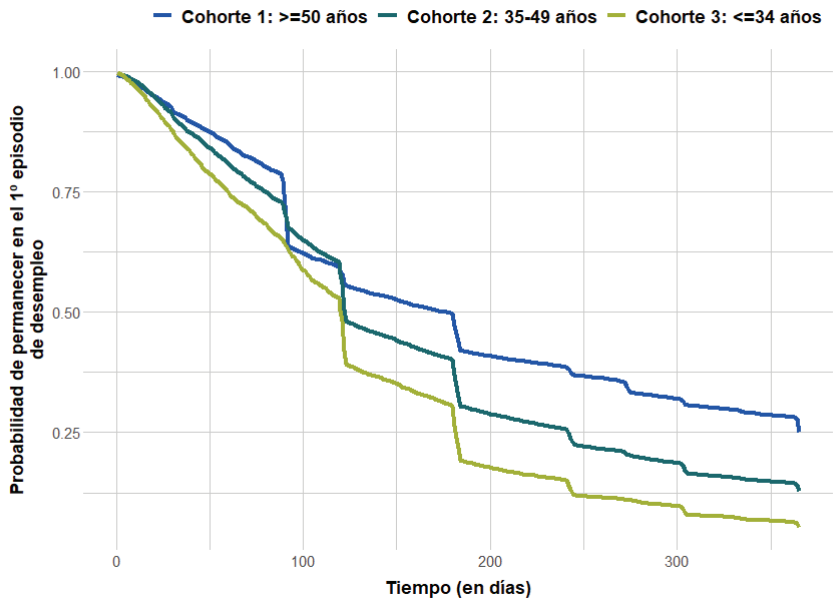
Tabla 5: Tabla de frecuencias de los tiempos (en número de días) de permanencia en el primer episodio de desempleo de los integrantes de la muestra longitudinal de la MCVL. Edición 2018, con la *estimación de Kaplan-Meier* de las funciones de supervivencia y riesgo, por cohortes de edad

ti	Cohorte 1:>=50 años		Cohorte 2: 35-49 años		Cohorte 3: <=34 años	
	Si	hi	Si	hi	Si	hi
1	0,9929	0,0071	0,9976	0,0024	0,9979	0,0021
2	0,9909	0,0021	0,9957	0,0019	0,9953	0,0027
3	0,9896	0,0013	0,9936	0,0020	0,9928	0,0025
4	0,9880	0,0017	0,9920	0,0016	0,9907	0,0021
5	0,9862	0,0018	0,9906	0,0014	0,9868	0,0040
6	0,9836	0,0027	0,9882	0,0025	0,9822	0,0046
…	…	…	…	…	…	…
28	0,9299	0,0035	0,9185	0,0037	0,8861	0,0053
29	0,9270	0,0031	0,9152	0,0036	0,8808	0,0060
30	0,9210	0,0065	0,9073	0,0086	0,8760	0,0054
31	0,9171	0,0042	0,9033	0,0045	0,8709	0,0059
32	0,9130	0,0045	0,8992	0,0045	0,8644	0,0074
…	…	…	…	…	…	…
179	0,4975	0,0017	0,4024	0,0035	0,3078	0,0026
180	0,4953	0,0045	0,4000	0,0061	0,3056	0,0072
181	0,4705	0,0501	0,3694	0,0764	0,2627	0,1403
182	0,4538	0,0355	0,3475	0,0592	0,2389	0,0904
183	0,4370	0,0369	0,3245	0,0661	0,2169	0,0924
184	0,4200	0,0389	0,3051	0,0600	0,1917	0,1159
…	…	…	…	…	…	…
363	0,2811	0,0016	0,1441	0,0049	0,0638	0,0102
364	0,2781	0,0110	0,1406	0,0246	0,0621	0,0274
365	0,2517	0,0948	0,1278	0,0907	0,0528	0,1489

Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

resultados (HU Ramón y Cajal, 2007; Sociedad Colombiana Cardiología y Cirugía Cardiovascular, 2017). En IGE (2022) se analiza extensamente qué diferencias existen entre utilizar una u otra fórmula; por cuestiones de espacio, esta comparativa no se va a reproducir en esta ponencia. Cuando hay observaciones censuradas es más correcto emplear el método de cálculo 1, por lo que, en lo sucesivo, estas van a ser las fórmulas empleadas.

Figura 7. *Estimación de Kaplan-Meier* de la función de supervivencia en el primer episodio de desempleo de los integrantes de la muestra longitudinal de la MCVL, por cohortes de edad



Fuente: elaboración propia a partir de la base de datos longitudinal construida empleando la MCVL. Edición 2018

4. CONCLUSIONES

En esta ponencia se ha analizado cómo ha evolucionado la inserción de la población en el mercado laboral gallego desde 1960 a la actualidad. Para realizar este estudio, nos hemos servido de los microdatos retrospectivos de la Muestra Continua de Vidas Laborales (MCVL) de la Seguridad Social, que permiten seguir la evolución laboral de las personas que la integran durante largos periodos de tiempo. Son muchas las fuentes que aportan información sobre esta temática, pero la mayor parte de ellas lo hacen desde una óptica transversal. Sólo la MCVL ha permitido, hasta el momento, poner en práctica un verdadero estudio de tipo longitudinal. Hemos obtenido resultados tan interesantes y, hasta el momento, tan poco cuantificados, como que la edad media de incorporación al empleo se ha reducido en las últimas décadas o que nuestros mayores disfrutaron de contratos más estables en su acceso al mercado laboral.

La aplicación de las técnicas del análisis de supervivencia a los datos longitudinales de la MCVL nos ha brindado una perspectiva diferente y complementaria que enriquece el estudio de la inserción laboral. Por medio de estas técnicas hemos comprobado que la probabilidad de firmar contratos de larga duración ha disminuido desde los años 70. No obstante, y como contraprestación a esta "precarización" de la inserción laboral, parece más factible salir de la situación de desempleo, al menos en lo que al primer episodio de este tipo se refiere: los jóvenes gallegos experimentan hoy con menor edad situaciones de falta de empleo pero, y seguramente porque se enfrentan a ellas antes, les cuesta menos tiempo encontrar un nuevo trabajo.

Los resultados presentados en esta ponencia son sólo una muestra de las posibilidades que ofrece este nuevo enfoque de estudio longitudinal del mercado laboral. En el IGE hemos puesto a disposición de nuestros usuarios y usuarias una parte de los resultados presentados en esta ponencia; en un futuro próximo esperamos seguir nutriendo nuestro conocimiento del mercado laboral con nuevos estudios realizados a partir de la base de datos longitudinal de la MCVL.

REFERENCIAS

Instituto Galego de Estatística (2022), *Estudo sobre a posibilidade de difundir información da Mostra Continua de Vidas Laborais (MCVL) con perspectiva lonxitudinal*, bajo petición expresa, empleando el siguiente formulario web:

https://www.ige.gal/web/peticioninfo.jsp?idioma=gl

Ministerio de Inclusión, Seguridad Social y Migraciones (2020), *MCVL. Muestra Continua de Vidas Laborales. Guía del contenido*, en red:

https://www.seg-social.es/wps/portal/wss/internet/EstadisticasPresupuestosEstudios/Estadisticas/EST211/1429

Instituto Galego de Estatística (2019), *Estudo da representatividade transversal da Mostra Continua de Vidas Laborais para o mercado laboral galego*, bajo petición expresa, empleando el siguiente formulario web:

https://www.ige.gal/web/peticioninfo.jsp?idioma=gl

Alonso Domínguez, Ángel (2018), *Análisis de la permanencia en el empleo de los trabajadores españoles durante el periodo 2007-2010*, en Papers. Revista de Sociología, volumen 103, número 3, año 2018, en red:

https://papers.uab.cat/article/view/v103-n3-alonso

Bentolilla, Samuel; García-Pérez, J. Ignacio; Jansen, Marcel (2018), *El paro de larga duración de los mayores de 45 años*, en Papeles de economía española, año 2018, número 156, en red:

https://dialnet.unirioja.es/ejemplar/494462

Veiguela Fernández, Noa; Romero Martínez, Pilar (2014), *Procesamiento y depuración de la Muestra Continua de Vidas Laborales (MCVL) para el estudio del mercado laboral gallego*, ponencia presentada en las XVIII Jornadas de Estadística de las Comunidades Autónomas que tuvieron lugar en Oviedo, del 3 al 4 de julio de 2014, bajo petición expresa, empleando el siguiente formulario web:

https://www.ige.gal/web/peticioninfo.jsp?idioma=gl

Recursos formativos sobre el análisis de supervivencia y su aplicación con R:

Therneau, T. (2023), *A Package for Survival Analysis in R. R package version 3.5-3*, en:

https://CRAN.R-project.org/package=survival

Kassambara, A.; Kosinski, M.; Biecek, P. (2021), *Survminer: Drawing Survival Curves using "ggplo2". R package version 0.4.9*, en:

https://CRAN.R-project.org/package=survminer

Urdinez, Francisco; Cruz Labrín, Andrés (editores) (2021), *AnalizaR Datos Políticos*, en:

https://arcruz0.github.io/libroadp/surv.html#el-modelo-cox-de-riesgos-proporcionales

Lastra, Iago (2019), *Análisis de supervivencia: Curvas de Kaplan-Meier en R*, en:

https://iagolast.github.io/blog/2019/01/13/kaplan-meier.html

Castro Kuriss, Mg. Claudia (2018), *Análisis de supervivencia mediante el empleo de R. Análisis de tiempos hasta un evento*, curso de Análisis de tiempos hasta un evento, impartido por el Instituto de Cálculo de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, en red:

https://www.researchgate.net/profile/Claudia-Castro-Kuriss/publication/325393091_Analisis_de_Sobrevida_mediante_el_software_R/links/5b0b200eaca2725783ea55c4/Analisis-de-Sobrevida-mediante-el-software-R.pdf

Martínez, Javier (2017), *Análisis de Supervivencia en R*, en:

https://rpubs.com/JavierMtzG/Supervivencia

Boj del Val, Eva (2017), *El modelo de regresión de Cox*, material formativo del Departamento de Matemática Económica, Financiera y Actuarial de la Facultad de Economía y Empresa de la Universidad de Barcelona, en red:

https://diposit.ub.edu/dspace/bitstream/2445/49070/6/El%20modelo%20de%20Cox%20de%20riesgos%20proporcionales.pdf

Materiales formativos sobre el análisis de supervivencia disponibles en la web de la Sociedad Colombiana de Cardiología y Cirugía Cardiovascular (2017), *Análisis de supervivencia*, en:

https://scc.org.co/wp-content/uploads/2017/10/Supervivencia.pdf

Hospital Universitario Ramón y Cajal (2007), *Análisis de supervivencia*, forma parte del material desarrollado en el marco de una conferencia pronunciada por Víctor Abraira en el 9º Congreso de la Sociedad Catalana del Transplante, del 25 al 28 de febrero de 2007, Barcelona, en red:

hrc.es/bioest/Supervivencia_1.html

Materiales formativos sobre el análisis de supervivencia facilitados por la Universidad de Santiago de Compostela (USC) en colaboración con la Fundación Pública Escola Galega de Administración Sanitaria (FEGAS), en:

https://www.usc.es/export9/sites/webinstitucional/gl/investigacion/grupos/psicom/docencia/grado/Modelos/Teoria/tema_6.pdf

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Estimação de momentos e densidade de tempos de primeira passagem por limiares de risco

Nuno M. Brites[1]

[1]ISEG - Instituto Superior de Economia e Gestão, Universidade de Lisboa; REM - Research in Economics and Mathematics, CEMAPRE, Rua do Quelhas, 6, gabinete 503, 1200 - 781 Lisboa

RESUMO

As equações diferenciais estocásticas podem ser usadas para explicar a dinâmica do tamanho de uma população em ambiente aleatório. Com base em expressões gerais para a média e desvio padrão dos tempos de primeira passagem por limiares de risco, calculamos esses valores para o caso específico do modelo logístico com pesca, tendo em consideração vários valores de limiares de risco. Mostramos ainda como, para um dado limiar de risco, a função de densidade de probabilidade do tempo para atingir os limiares de risco pode ser estimada invertendo numericamente a transformada de Laplace.

Palavras e frases chave: Limiar de risco, tempos de primeira passagem, equações diferenciais estocásticas, modelo logístico

REFERÊNCIAS

Brites, N. M. (2022) Moments and probability density of threshold crossing times for populations in random environments under sustainable harvesting policies. Computational Statistics.

Brites, N. M. (2022) Fisheries management in randomly varying environments: Comparison of constant, variable and penalized efforts policies for the Gompertz model. Fisheries Research, 216, 196–49.

Giet, J. S., Vallois P., Wantz-Mézières, S. (2015) The logistic SDE. Proceedings. Theory of Stochastic Processes, 20(36), 28–62.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

REVISITING COMPUTATIONAL PROCEDURES FOR IMPROVING PARAMETER ESTIMATION IN EXTREMES

M. Manuela Neves[1]
manela@isa.ulisboa.pt

[1] CEAUL & Instituto Superior de Agronomia, Universidade de Lisboa, Portugal

ABSTRACT

Accurate estimation of tail distributions is very important in areas where extremes or catastrophic events may occur. The main difficulty from the statistical perspective is that the available data to base the estimates on is very sparse, which calls for tailored estimation methods. When modelling extreme events there are a few primordial parameters among which we refer to the *extreme value index* (EVI), denoted by ξ, and the *extremal index* (EI), denoted by θ. The EVI measures the right tail-weight of the underlying distribution and the EI characterizes the degree of local dependence in the extremes of a stationary sequence. Most of the semi-parametric estimators of these parameters present the well known type of behaviour: nice asymptotic properties but a high variance for small k, the number of upper order statistics used in the estimation, and an increasing bias with k. Recently, computer intensive procedures have revealed to be highly fruitful in extreme value parameter estimation. The role of computer intensive methodologies jointly with adaptive algorithms for an adequate estimation of the aforementioned parameters are here revisited.

Keywords: adaptive algorithm, computational procedures, extremal index, extreme value index, semi-parametric estimation.

1. INTRODUCTION

Rare events are part of the real world, but extreme environmental events can have a huge impact on everyday life. We are familiar, for example, with the consequences and damage caused by hurricanes, floods and other major natural disasters. Consequently, there has been considerable attention to studying, understanding and predicting the nature of such phenomena and the problems caused by them. The consideration of the major risks in our technological society has become vital because of the economic, environmental and human impacts of disasters. One of the standard approaches to studying risks is Extreme Value Theory (EVT), a branch of statistics dealing with the extreme deviations from the median of probability distributions.

The study of extreme events has become increasingly important, both in terms of probabilistic and statistical research. EVT aims to study and to predict the occurrence of extreme or even rare events, outside of the range of available data.

As already said the main objective of the EVT is to know or predict the statistical probabilities of events that have never (or rarely) been observed. Firstly, the statistical analysis of extreme values has been developed in order to study flood levels. Nowadays, the domains of application include other meteorological events (such as precipitation or wind speed), industry (for example important malfunctions), finance (e.g. financial crises), insurance (for very large claims due to catastrophic events), environmental sciences (like concentration of ozone in the air).

Extreme Value distributions arise as limiting distributions for maximums or minimums (extreme values) of a sample of independent, identically distributed (i.i.d.) random variables, as the sample size increases. EVT is the theory of modelling and measuring events which occur with very small probability.

The Generalized Extreme Value distribution, defined below, was first due to Fréchet (1927), Fisher and Tippett (1928), Gumbel (1935) and von Mises (1936). But were Gnedenko (1943) and de Haan(1970) who gave conditions for its complete characterization.

The classical assumption in EVT is that we have a set of i.i.d. random variables (r.v.'s), $X_1, \ldots, X_n$, from an unknown cumulative distribution function (c.d.f.) F and we are concerned with the limit behaviour of $M_n \equiv X_{n:n} = \max(X_1, \ldots, X_n)$ as $n \to \infty$.

Whenever it is possible to linearly normalize M_n so that we get a non-degenerate limit, as $n \to \infty$, such a limit is of the type of the extreme value (EV) d.f.,

$$\mathrm{EV}_\xi(x) := \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0 \quad \text{if} \quad \xi \neq 0 \\ \exp[-\exp(-x)], & x \in \Re \qquad\quad \text{if} \quad \xi = 0. \end{cases} \tag{1}$$

We then say that F is in the domain of attraction for maxima of EV_ξ, denoting this by $F \in D_{\mathcal{M}}(\mathrm{EV}_\xi)$.

However (1) can also incorporate location (λ) and scale ($\delta > 0$) parameters, and in this case, the EV_ξ d.f. is given by,

$$\mathrm{EV}_\xi(x; \lambda, \delta) \equiv \mathrm{EV}_\xi((x-\lambda)/\delta).$$

The parameter ξ is the *extreme value index* (EVI) and it measures essentially the weight of the right tail function, $\overline{F} = 1 - F$. The parameter ξ is also the basis of other important parameters of extreme events, such as:

- a *high quantile* of probability $1-p$ (p small)

$$\begin{aligned} \chi_{1-p} &:= \inf\{x : F(x) \geq 1-p\}, \\ \chi_{1-p} &:= \lambda - \frac{\delta}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], \quad \xi \neq 0 \end{aligned}$$

- the *probability of exceedance* of a high level;

- the *return period* of a high level,

- the *right endpoint* of an underlying model F,

$$w_F := \{x \in \mathbb{R} : F(x) < 1\}.$$

The estimation of ξ, in (1), is then of primordial importance not only by itself but also because it is the basis for the estimation of all other parameters of extreme events.

In most fields of applications, the independence assumption is not valid. Stationary sequences are realistic for many real problems and dependence in stationary sequences can assume several forms. Provided that a stationary sequence $\{X_n\}_{n\geq 1}$ has limited long-range dependence at extreme levels, the maxima of this sequence follow the same distributional limit law as the associated independent sequence, $\{Y_n\}_{n\geq 1}$, but with other values for the parameters of EV d.f., Leadbetter *et al.*, (1983). Let us assume to be working with a strictly stationary sequence of r.v.'s, $\{X_n\}_{n\geq 1}$, with marginal d.f. F, under the long range dependence condition **D** (Leadbetter *et al.*, 1983) and the local dependence condition $\mathbf{D}''$ (Leadbetter and Nandagopalan, 1989). The stationary sequence $\{X_n\}_{n\geq 1}$ is said to have an extremal index (EI), θ, $0 < \theta \leq 1$, if for each $\tau > 0$, we can find a sequence of levels $u_n = u_n(\tau)$ such that, with $\{Y_n\}_{n\geq 1}$ the associated i.i.d. sequence (i.e. from the same F),

$$\mathrm{P}\left(Y_{n:n} \leq u_n\right) = F^n(u_n) \underset{n\to\infty}{\longrightarrow} e^{-\tau} \quad \text{and} \quad \mathrm{P}\left(X_{n:n} \leq u_n\right) \underset{n\to\infty}{\longrightarrow} e^{-\theta\tau}.$$

Under the validity of those conditions the extremal index can also be defined as

$$\theta = \frac{1}{\text{limiting mean size of clusters}} = \lim_{n\to\infty} \mathrm{P}(X_2 \leq u_n | X_1 > u_n) = \lim_{n\to\infty} \mathrm{P}(X_2 \geq u_n | X_1 < u_n), \tag{2}$$

where $u_n :\ F(u_n) = 1 - \tau/n + o(1/n)$, as $n \to \infty$, with $\tau > 0$, fixed.

D and $\mathbf{D}''$ are straightforwardly valid for i.i.d. data and $\theta = 1$.

2. EVI AND EI ESTIMATION: A FEW DETAILS

Under a semi-parametric framework it is only necessary to assume that $F \in D_{\mathcal{M}}(\mathrm{EV}_\xi)$ and the estimators of EVI are based on the k largest observations. Considering heavy-tailed parents, the sample $(X_1, X_2, \cdots, X_n)$ and the associated sample of ascending order statistics (o.s.'s), $(X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n})$, the most famous classical EVI-estimator is the Hill estimator (Hill, 1975), here denoted $H(k)$ and given by

$$H(k) := \frac{1}{k} \sum_{i=1}^{k} \{\ln X_{n-i+1:n} - \ln X_{n-k:n}\}, \qquad k = 1, 2, \ldots, n-1. \tag{3}$$

Consistency of (3) is achieved if $X_{n-k:n}$ is an *intermediate* o.s., i.e., if

$$k \equiv k_n \to \infty \quad \text{and} \quad k/n \to 0, \text{ as } n \to \infty,$$

but that estimator presents a very high variance for small values of k and a very high bias for large values of k. A simple class of second-order *minimum-variance reduced-bias* (MVRB) EVI-estimators is the one in Caeiro *et al.* (2005). This class, here denoted $\overline{H}(k)$, depends upon the estimation of second-order parameters (β, ρ) and has the functional form:

$$\overline{H}(k) := H(k)\big(1 - \widehat{\beta}(n/k)^{\widehat{\rho}}/(1-\widehat{\rho})\big),$$

with $H(k)$ the Hill estimator in (3), and where $(\widehat{\beta}, \widehat{\rho})$ needs to be an adequate consistent estimator of (β, ρ), as given, for example, in Gomes and Pestana (2007).

Concerning the θ parameter, the classical up-crossing, UC-estimator, $\widehat{\Theta}^{UC}$ (Nandagopalan, 1990), is a naive estimator that comes directly as an empirical counterpart of (2),

$$\widehat{\Theta}^{UC}(u_n) := \frac{\sum_{i=1}^{n-1} I\left(X_i \leq u_n < X_{i+1}\right)}{\sum_{i=1}^{n} I(X_i > u_n)},$$

for a suitable threshold u_n, where $I(A)$ denotes, as usual, the indicator function of A. Consistency of this estimator is obtained provided that the high level u_n is a normalized level, i.e. if with $\tau \equiv \tau_n$ fixed, the underlying d.f. F verifies $F(u_n) = 1 - \tau/n + o(1/n), \quad n \to \infty$ and $\tau/n \to 0$.

Regarding the θ estimation and considering u_n as a deterministic level $u \in [X_{n-k:n}, X_{n-k+1:n}[$, that estimator can be written as (see Gomes *et al.*, 2008)

$$\widehat{\Theta}^{UC}(k) := \frac{\sum_{i=1}^{n-1} I\left(X_i \leq X_{n-k:n} < X_{i+1}\right)}{k}. \tag{4}$$

However this estimator shows the same drawbacks already mentioned.

3. COMPUTATIONAL RESOURCES

Kleinow and Thomas (2000) present a brief review of some computational resources for extremes, describing and showing examples of statistical software systems.

Gilleland *et.al* (2013) have an excellent review of **R** *packages*, R Core Team (2022), for extreme value analysis. Among other packages in **R** for EVT analysis we can mention `evd`, `ismev`, `fExtremes`, `evir`, `extRemes`, `evdbayes`, `extremevalues`, `copula` and `SpatialExtremes`. They discuss and explain the main packages and also compare to other ones existing in other softwares such as MatLab or even packages written in C++ or Fortran.

Among those computational facilities we can also mention some computational resampling procedures such as Jackknife and Bootstrap (Efron and Tibshirani, 1994). They have revealed to give good results in the reduction of the bias of an estimator as well as in the improvement of the threshold selection.

Gomes *et al.* (2013) considered to remove the non-null asymptotic bias of the EVI-estimators through the use of the Generalized Jackknife (GJ) methodology (see Gray and Schucany (1972)), by using an adequate pair of EVI-estimators to build a reduced-bias affine combination of them.

As an example, we can refer to a class of GJ–EVI estimators, parameterised in a tuning parameter $\alpha \in (0,1)$, defined as

$$\overline{H}^{GJ}(k) := \frac{\overline{H}(k) - \alpha^{2\hat{\rho}}\overline{H}(\lfloor \alpha k \rfloor)}{1-\alpha^{2\hat{\rho}}},$$

where $\lfloor x \rfloor$ denotes, as usual, the integer part of x. Only as an illustration we will show an application of this estimator with $\alpha = 1/2$.

Gomes *et al.* (2008) considered the estimator for θ defined in (4) and derived a reduced-bias GJ estimator of order 2, based on the estimator $\widehat{\Theta}^{UC}$ computed at the three levels, k, $\lfloor k/2 \rfloor + 1$ and $\lfloor k/4 \rfloor + 1$, given by

$$\widehat{\Theta}_n^{GJ} := 5\widehat{\Theta}^{UC}(\lfloor k/2 \rfloor + 1) - 2\big(\widehat{\Theta}^{UC}(\lfloor k/4 \rfloor + 1) + \widehat{\Theta}^{UC}(k)\big). \tag{5}$$

$\widehat{\Theta}_n^{GJ}$ estimator, although presenting a sample path around the true parameter value, shows a high volatility. An appropriate choice of level k is therefore highly necessary and a problem not yet completely resolved.

Adaptive choice of threshold k, through the bootstrap methodology and also through some heuristic computational procedures have been studied in several works, such as Gomes *et al.* (2012), Neves *et al.* (2015) and Caeiro and Gomes (2016), to cite only a few works.

One of the adaptive procedures is detailed in the following algorithm

Algorithm: Let us generally denote by $T(k)$ any of the above mentioned estimators for ξ or θ.

Step 1. Given an observed sample $(x_1, \ldots, x_n)$, compute, for $k = 1, \ldots, n-1$, the observed values of $T(k)$.

Step 2. Obtain j_0, the minimum value of j, a non-negative integer, such that the rounded values, to j decimal places, of the estimates in Step **1** are distinct. Define $a_k^{(T)}(j) = \text{round}(T(k), j)$, $k = 1, 2, \ldots, n-1$, the rounded values of $T(k)$ to j decimal places.

Step 3. Consider the sets of k values associated to equal consecutive values of $a_k^{(T)}(j_0)$, obtained in Step **2.** Set $k_{min}^{(T)}$ and $k_{max}^{(T)}$ the minimum and maximum values, respectively, of the set with the largest range. The largest run size is then $l_{_T} := k_{max}^{(T)} - k_{min}^{(T)}$.

Step 4. Consider all those estimates, $T(k)$, $k_{min}^{(T)} \le k \le k_{max}^{(T)}$, now with two extra decimal places, i.e. compute $T(k) = a_k^{(T)}(j_0 + 2)$. Obtain the mode of $T(k)$ and denote $\mathcal{K}_{_T}$ the set of k-values associated with this mode.

Step 5. Take $\hat{k}_{_T}$ as the maximum value of $\mathcal{K}_{_T}$, and consider the adaptive estimate $T(\hat{k}_{_T})$.

Step 6. The best estimate is the value of T that corresponds to the maximum run size $l_{_T}$ computed in Step **3**.

To investigate the performace of this algorithm a Monte Carlo simulation was performed in Neves *et al.* (2015) and here briefly illustrated in the next section.

Boostrap methodololy performed here by resampling blocks of observations instead of single observations, is now a topic under research. A few results have already been presented, Prata Gomes and Neves (2015 a,b) and Neves (2015).

Recently, Ferreira (2023) considered the estimator proposed by Ferreira and Ferreira (2018), on which Jackknife and Bootstrap resampling techniques were applied for bias reduction and interval estimation.

4. MONTE CARLO SIMULATIONS: AN ILLUSTRATION

Monte Carlo simulation studies allow to investigate the performance of our procedure. Here two models illustrate the behaviour of those computational methods: the Fréchet model for an i.i.d. setup and the ARMAX model:

The Fréchet model

Let $\{X_i\}_{i\geq 1}$ be a sequence of independent random variables, with d.f. $F(x) = \exp(-x^{-1/\xi})$, $x > 0$. Here we have then $\theta = 1$.

Max-Autoregressive Process

Let $\{Z_i\}_{i\geq 1}$ be a sequence of independent, unit-Fréchet distributed random variables. For $0 < \theta \leq 1$, let $Y_1 = Z_1$ and $Y_i = \max\{(1-\theta)Y_{i-1}, \theta Z_i\}$, for $i \geq 2$.

The marginal distribution of the process $\{Y_i\}_{i\geq 1}$ is unit-Fréchet and for $u_n = ny$, $0 < y < \infty$, $P\{M_n \leq u_n\} \to \exp\big(-\theta/y\big)$, as $n \to \infty$.

The *extremal index* of the sequence is equal to θ (see Beirlant *et al.*, 2004).

Samples of size 1000 were generated for some values of the unknown parameters in the two mentioned models.

Figures 1 and 2 show, for each model, the values of the parameters considered, the sample paths of the estimates and the results obtained from the application of the adaptive algorithm.

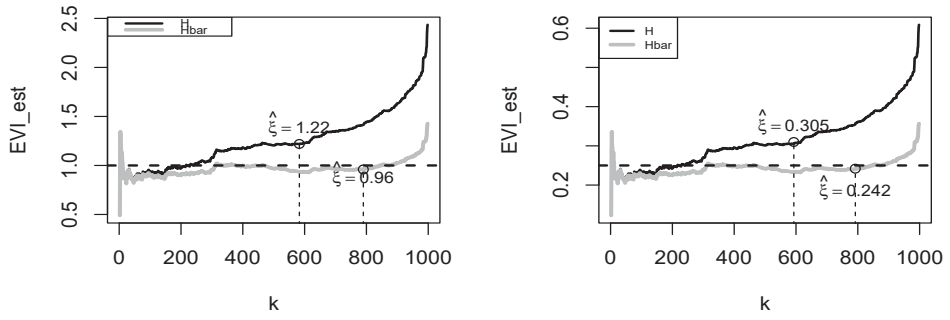


Figure 1: Adaptive choice of the level k for estimating ξ for random samples from a Fréchet model, with $\xi = 1$ (left) and $\xi = 0.25$ (right).

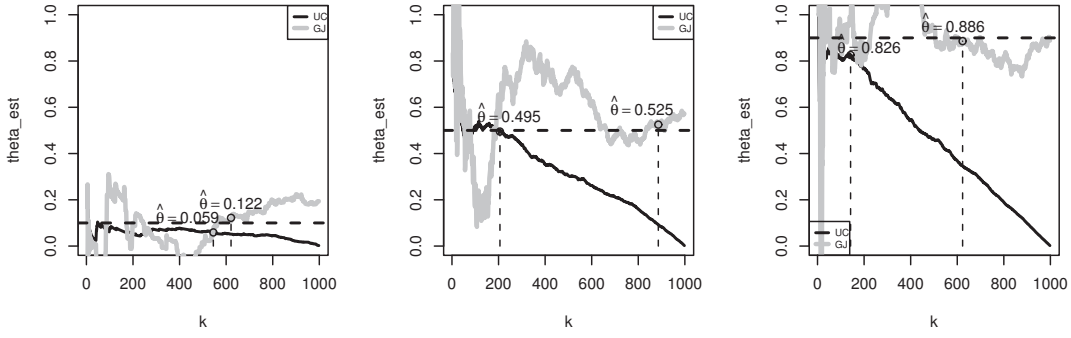


Figure 2: Adaptive choice of the level k for estimating θ for random samples generated with $\theta = 0.1, 0.5, 0.9$ (left to right) from the ARMAX model.

5. CONCLUDING NOTES AND WORK IN PROGRESS

The heuristic algorithm described above seems to perform very well in the choice of the level k to be used in the estimation of both ξ and θ. Further simulation studies are now in progress. For

the estimation of ξ, and as expected, the algorithm leads to large k-estimates and consequently to more reliable estimates of ξ. A similar comment applies to the adaptive estimation of θ.

However we think that future research is still welcome for improving the estimators of θ, so that more stable patterns can be obtained and possibly other alternative adaptive estimators too. Applications to real data will also be performed.

Acknowledgements

This work is partially financed by national funds through FCT –Fundação para a Ciência e a Tecnologia through the project UIDB/00006/2020 (CEAUL).

REFERENCES

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D. and Ferro, C. (2004) Statistics of Extremes: Theory and Applications. John Wiley & Sons.

Caeiro, F. and Gomes, M.I. (2016) Threshold Selection in Extreme Value Analysis. Extreme Value Modeling and Risk Analysis. 69 – 86. Chapman and Hall

Caeiro, F., Gomes, M.I. and Pestana, D.D. (2005) Direct reduction of bias of the classical Hill estimator, Revstat 3:2 111–136.

de Haan, L. (1970) On Regular Variation and its Applications to the Weak Convergence of Sample Extremes, Mathematical Centre Tract 32, Amesterdam, Dordrecht: D.Reidel.

Efron, B. and Tibshirani, R. (1994) An introduction to the Bootstrap. Chapman & Hall.

Ferreira, M. (2023) Extremal index: estimation and resampling. Computational Statistics https://doi.org/10.1007/s00180-023-01406-91 3

Ferreira, H. and Ferreira M. (2018) Estimating the extremal index through local dependence. Ann Inst Henri Poincaré Probab Stat 54(2):587-605

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24, 180–190.

Fréchet, M. (1927) Sur la loi de probabilité de l'écart maximum, Ann. Soc. Polon. Math. (Cracovie), 6, 93–116 .

Gilleland, E., Ribatet, M. and Stephenson, A.G. (2013) A software review for extreme value analysis, Extremes 16: 103–119

Gomes M.I. and Pestana D. (2007) A sturdy reduced-bias extreme quantile (VaR) estimator. J. Amer. Statist. Assoc., 102 280–292.

Gomes, M.I., Figueiredo, F. and Neves, M.M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action. *Extremes* **15**, 463–489.

Gomes, M.I., Hall, A. and Miranda, C. (2008) Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. J. Comput. Statist. and Data Analysis 52:4, 2022–2041.

Gomes, M. I., Martins, M. J. and Neves, M. M. (2013) Generalised Jackknife-based estimators for univariate extreme-value modeling. Comm. Statist. Theory Methods 42:7 1227–1245.

Gnedenko, B. V. (1943) Sur la distribution limite d'une série aléatoire. Annals of Mathematics, 44, 423–453.

Gray, H.L. and Schucany, W.R. (1972) The Generalized Jackknife Statistic. Marcel Dekker. New York.

Gumbel, E. J. (1935) Les valeurs extrêmes des distributions statistiques. Ann. Inst. Henri Poincaré, 5, 2, 115–158.

Hill, H. (1975) A simple general approach to inference about the tail of a distributio. Ann. Statist., 1163–1174.

Kleinow, T. and Thomas, M. (2000) Computational Resources for Extremes. In: Franke, J., Stahl, G., Härdle, W. (eds) Measuring Risk in Complex Stochastic Systems. Lecture Notes in Statistics, vol 147. Springer, New York.

Leadbetter, M. R., Lindgren G. and Rootzén H. (1983) Extremes and related properties of random sequences and series. Springer-Verlag, New York.

Leadbetter, M.R. and Nandagopalan, L. (1989) On exceedance point process for stationary sequences under mild oscillation restrictions. Extreme Value Theory: Proceedings, Oberwolfach J. Hüsler and R.D. Reiss (eds.), Lecture Notes in Statistics 52, 69–80. Springer-Verlag, Berlim.

Nandagopalan, S. (1990) Multivariate Extremes and Estimation of the Extremal Index, PhD Thesis, University of North Carolina, Chapel Hill.

Neves, M. M. (2015) Bootstrap and Jackknife methods in extremal index estimation: A review. in E. Gonçalves, P. Oliveira, P. and C. Tenreiro (eds.) Contributions in Statistics and Inference: Celebrating Nazaré Mendes Lopes' Birthday, Textos de Matemática,DMUC, 47, 49–66.

Neves, M.M., Gomes, M.I., Figueiredo, F. and Prata Gomes, D. (2015) Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation. J. Stat. Theory Pract., 9:1, 184–199.

Prata Gomes, D. and Neves, M.M. (2015 a) Bootstrap and Other Resampling Methodologies in Statistics of Extremes. Communications in Statistics–Simulation and Computation, 44:10, 2592-2607.

Prata Gomes, D. and Neves, M. M. (2015 b) Adaptive choice and resampling techniques in extremal index estimation. In Kitsos, C., Oliveira, T., Rigas, A. and Gulati, S. (eds.), Theory and Practice of Risk Assessment, Springer Proceedings in Mathematics and Statistics, 321-332.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

von Mises, R. (1936) La distribution de la plus grande de n valeurs. American Mathematical Society, Reprinted in Selected Papers Volumen II, Providence, R.I., 271–294.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

AUC optimism correction in logistic regression with missing data

Susana Rafaela Guimarães Martins[1], Maria del Carmen Iglesias-Pérez[2] and Jacobo de Uña-Álvarez[3]

[1] Escola Superior de Desporto e Lazer, Instituto Politécnico de Viana do Castelo; CINBIO, Universidade de Vigo
[2] Department of Statistics and OR, Universidade de Vigo; CINBIO, Universidade de Vigo
[3] Department of Statistics and OR, Universidade de Vigo; CINBIO, Universidade de Vigo

ABSTRACT

Logistic regression is a well-known approach to predict a binary outcome given covariates. To evaluate the predictive capacity of a regression model, the Area Under the Curve (AUC) is often used. Note that when the same sample is used to fit the model and estimate its predictive ability, there may be optimism in the AUC. The AUC estimated in this way is called the apparent AUC and some correction methods have been studied in the complete data context.

In this work we investigate the issue of estimating the AUC in the presence of missing data in the covariables, that are continuous. For the construction of the predictive models, different missing data methodologies were applied: Complete Case Analysis, Inverse Probability Weighting and Multiple Imputation, and the apparent AUC was estimated for each one of them. With a simulation study we evaluate the performance of the several estimators for the AUC; in particular, the Monte Carlo bias and mean squared error of the estimators are obtained. The bias is defined as the difference between apparent AUC and out-of-sample AUC, where out-of-sample AUC approximates the true AUC, that is, the AUC of the population.

Traditionally, the apparent AUC overestimates the true AUC. In this work we consider several approaches to correct for this overestimation: split-sample, k-fold and leave-one-out, adapted to missing data. We consider different missing scenarios: missing completely at random, missing at random and missing not at random. In the simulation study we also evaluate the performance of the correction methods in the presence of missing data.

Keywords: Missing data, AUC optimism correction, prediction, ROC curve

REFERENCES

Iparraguirre, A., Irantzu, B., Rodríguez-Álvarez, M.X. (2019) On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. SORT-Statistics and Operations Research Transactions, 1, 145-162, 2019.

Li, P., Taylor, J.M.G., Spratt, D.E., Karnes, R.J., Schipper, M.J. (2021) Evaluation of predictive model performance of an existing model in the presence of missing data. Statistics in Medicine, 40, 3477–3498.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Density regression via Dirichlet process mixtures of normal structured additive regression models

María Xosé Rodríguez Álvarez[1,2], Vanda Inácio[3]

[1]CINBIO, Universidade de Vigo, Department of Statistics and Operations Research, Vigo, Spain
[2]CITMAga, Galician Center for Mathematical Research and Technology, Santiago de Compostela, Spain
[3]School of Mathematics, University of Edinburgh, Scotland, United Kingdom

ABSTRACT

In many real-life applications, it is of interest to study how the distribution of a univariate, real-valued, continuous response changes with a set of covariates. Within a Bayesian nonparametric framework, dependent Dirichlet process mixture of normal distributions provide a highly flexible approach for estimating the conditional density function (De Iorio et al., 2009; Quintana et al., 2022). However, several formulations of this class of models involve intricate algorithms for posterior inference, thus preventing their widespread. Motivated by this problem, we propose a flexible, versatile, and computationally tractable model for density regression based on a dependent Dirichlet process mixture of normals model where an additive structure is assumed for the mean of each component and the effects of continuous covariates are modelled trough smooth functions. The major modelling components are penalised B-splines (P-splines; Eilers and Marx, 1996; Lang and Brezger, 2004) and their bivariate tensor product extension. The resulting model can be regarded as a Dirichlet process mixture of normal structured additive regression models and can easily deal with discrete covariates, nonlinear effects of continuous covariates, interaction surfaces, spatial effects, and varying coefficient terms. We coin our approach as DDPstar. A practically important feature of DDPstar models is that all parameters have conjugate full conditional distributions thus leading to straightforward Gibbs sampling. The validity of our approach is supported by simulations and applied to a real dataset concerning the study of the association of a toxic metabolite on preterm birth.

Keywords: Bivariate smoothing; density regression; dependent Dirichlet process; Gibbs sampling; P-splines; structured additive regression.

REFERENCES

De Iorio, M., Johnson, W.O., Müller, P. and Rosner, G.L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. Biometrics, 65, 762–771.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11, 89–121.

Quintana, F.A., Müller, P., Jara, A., and MacEachern, S.N. (2022). The dependent Dirichlet process and related models. Statistical Science, 37, 24–41.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. Journal of Computational and Graphical Statistics, 13, 183–212.

Técnicas de reducción de dominio en optimización no lineal

Ignacio Gómez-Casares[1,2], Julio González Díaz[1,2], Brais González Rodríguez[3] y Pablo Rodríguez-Fernández[2]

[1] CITMAga (Centro de Investigación y Tecnología Matemática de Galicia)
[2] Universidad de Santiago de Compostela
[3] Ivey Bussiness School

RESUMEN

Una de las principales formas de mejorar el rendimiento de un esquema de ramificación y acotación es la reducción del tamaño del árbol. Las técnicas de ajuste de cotas se aplican normalmente en el contexto de problemas MILP y, aquí, extendemos varias de esas técnicas al contexto de un algoritmo de ramificación y acotación no lineal. También incluimos nuevas formas de realizar el ajuste de cotas mediante el uso de restricciones SDP y SOCP. Además del ajuste de cotas, la selección del punto de ramificación en cada nodo del árbol también tiene un gran impacto en el tamaño total del árbol. Analizamos el rendimiento de varias formas de elegir el punto de ramificación en un árbol no lineal. Por último, utilizamos un enfoque de selección de carteras mediante aprendizaje automático y elegimos la mejor configuración para un problema concreto.

Palabras y frases clave: optimización polinómica, machine learning, bound-tightening, spatial branching

REFERENCIAS

Pietro Belotti, Sonia Cafieri, Jon Lee, and Leo Liberti, On feasibility based bounds tightening, 2012.

Pietro Belotti, Jon Lee, Leo Liberti, François Margot, and Andreas Wächter, Branching and bounds tightening techniques for non-convex MINLP, Optimization Methods and Software 24 (2009oct), no. 4-5, 597–634.

Michael R. Bussieck, Arne Stolbjerg Drud, and Alexander Meeraus, MINLPLib-a collection of test models for mixed-integer nonlinear programming, INFORMS Journal on Computing 15 (2003), 114–119.

Evrim Dalkiran and Hanif D. Sherali, Theoretical filtering of RLT bound-factor constraints for solving polynomial programming problems to global optimality, Journal of Global Optimization 4 (2013), 1147–1172.

Evrim Dalkiran and Hanif D. Sherali, RLT-POS: Reformulation-linearization technique-based optimization software for solving polynomial programming problems, Mathematical Programming Computation 8 (2016), 337–375.

Fabio Furini, Emiliano Traversi, Pietro Belotti, Antonio Frangioni, Ambros Gleixner, Nick Gould, Leo Liberti, Andrea Lodi, Ruth Misener, Hans Mittelmann, Nikolaos Sahinidis, Stefan Vigerske, and Angelika Wiegele, QPLIB: a library of quadratic programming instances, Mathematical Programming Computation 1 (2018), 237–265.

Bissan Ghaddar, Ignacio Gómez-Casares, Julio González-Díaz, Brais González-Rodríguez, Beatriz Pateiro-López, and Sofía Rodríguez-Ballesteros, Learning for spatial branching: An algorithm selection approach, INFORMS Journal on Computing (2023may).

Ambros M. Gleixner, Timo Berthold, Benjamin Müller, and Stefan Weltge, Three enhancements for optimization-based bound tightening, Journal of Global Optimization 67 (2017), no. 4, 731–757.

Brais González-Rodríguez, Raúl Alvite-Pazó, Samuel Alvite-Pazó, Bissan Ghaddar, and Julio González-Díaz, Polynomial optimization: Enhancing RLT relaxations with conic constraints, 2023. arXiv.

Brais González-Rodríguez, Joaquín Ossorio-Castillo, Julio González-Díaz, Ángel M González-Rueda, David R Penas, and Diego Rodríguez-Martínez, Computational advances in polynomial optimization: RAPOSa, a freely available global solver, Journal of Global Optimization (2022).
Hong S. Ryoo and Nikolaos V. Sahinidis, A branch-and-reduce approach to global optimization, Journal of Global Optimization 8 (1996), no. 2, 107–138.
Hanif D. Sherali and Cihan H. Tuncbilek, A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique, Journal of Global Optimization 1 (1992), 101–112.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

ASIGNACIÓN ÓPTIMA DE MEDIOS AÉREOS NA EXTINCIÓN DUN GRAN INCENDIO FORESTAL: MODELO E HEURÍSTICA

Marta Rodríguez Barreiro[1,2], María José Ginzo Villamayor[2,3], Fernando Pérez Porras[4], María Luisa Carpente Rodríguez[1], Silvia María Lorenzo Freire[1,2,5]

[1] Universidade da Coruña, Departamento de Matemáticas.
[2] Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga).
[3] Universidade de Santiago de Compostela, Departamento de Estatística, Análise Matemática e Optimización.
[4] Universidad de Córdoba, Departamento de Enxeñería Gráfica e Xeomática.
[5] Centro de Investigación en Teconoloxías da Información e as Comunicacións (CITIC).

RESUMO

A lacra dos incendios forestais continúa a ser un dos grandes problemas deste século a nivel mundial. En Galicia o número de incendios ocorridos conseguiu reducirse nos últimos anos, pero a magnitude destes é maior polo que o número de hectáreas queimadas aumentou (Consellería de Medio Rural, 2023). Segundo a lexislación española cando nun incendio traballan 5 ou máis aeronaves de extinción é obrigatoria a presencia do coordinador de medios aéreos, cuxas principais funcións son a asignación das aeronaves ás zonas de extinción do incendio en cada tramo horario, a asignación dos puntos de carga de auga, e as bases de descanso entre outras. Todas estas decisións deben ser tomadas polo coordinador do incendio co único apoio dunha guía escrita, o que lle imposibilita automatizar esta labor. Neste traballo proponse a creación dun modelo de programación lineal enteira mixta para axudar nestas labores de coordinación de medios aéreos, que proporcione a planificación óptima das aeronaves no seu traballo de extinción dun gran incendio forestal, asignando as aeronaves que deben traballar en cada período de tempo e decidindo en que puntos do incendio deben actuar, en que puntos cargar auga e as bases de descanso asignadas a cada aeronave. Dado o contexto de emerxencia, a planificación debe estar dispoñible en poucos minutos para tomar decisións en tempo real que podan influír no control do incendio. Porén, dada a complexidade do modelo, tarda varias horas en resolverse empregando solvers como Gurobi en escenarios de moitas aeronaves. Para solucionar isto, proponse o desenvolvemento dunha heurística para acadar bos tempos de resolución.

Palabras e frases chave: MILP, heurística, simmulated annealing, asignación de recursos, incendios forestais

1. INTRODUCIÓN

A tendencia dos últimos anos en Galicia segundo o Plan de Prevención e Defensa contra os Incendios Forestais de Galicia (PLADIGA, Consellería de Medio Rural, 2023) é que cada vez hai menos incendios pero son máis grandes, con maior superficie afectada. Considérase que un incendio é un gran incendio cando este ten unha superficie maior a 500 hectáreas. Nestes grandes incendios, nos que traballan un gran número de efectivos para a súa extinción, é obrigatoria a presenza da figura do coordinador de medios aéreos.

O coordinador de medios aéreos actúa asignando as aeronaves dispoñibles ó incendio, tendo en conta a súa evolución esperada a partir da observación do terreo, do combustible e da predición meteorolóxica. Ás veces poden axudarse de simuladores de incendios para obter a evolución esperada do mesmo e seleccionar as zonas de ataque das aeronaves.

Na literatura existen moitos traballos relacionados coa asignación óptima de medios de extinción a un incendio forestal. En Granda *et al.* (2023) pode atoparse unha boa revisión bibliográfica de todos os modelos existentes na literatura que tratan de abordar este problema. Porén, non se atopou ningún modelo que trate de establecer a planificación completa de cada aeronave que traballará no incendio, tendo como obxectivo maximizar a eficacia das descargas de auga efectuadas, e decidindo tamén os puntos nos que as aeronaves cargarán e descargarán auga, así como as bases de descanso.

2. MODELO DE PLANIFICACIÓN DE AERONAVES DE EXTINCIÓN

O modelo de optimización desenvolto trátase dun modelo de programación enteira mixta (MILP) multiobxectivo que maximiza a eficiencia das descargas de auga das aeronaves penalizando os voos das mesmas, os cambios de traxectoria e non satisfacer os requisitos do coordinador do incendio. Minimizar os custos económicos non son o principal obxectivo deste modelo, a diferencia de moitos atopados na literatura que se basean no proposto por Donovan *and* Rideout (2003). Ademais, cumpre a normativa Circular Operativa 16-B (Dirección General de Aviación Civil, 1995) que regula as limitacións das actividades aéreas en España, baseado nun modelo proposto por Rodríguez-Veiga *et al.* (2018).

O modelo decide que aeronaves traballan no incendio e en que períodos de tempo o fan, en que puntos cargan e descargan auga as aeronaves, e en que bases deben descansar en cada período de tempo. No traballo proposto por Rodríguez-Veiga *et al.* (2018) abórdanse obxectivos similares, empregando dous modelos d programación lineal. A principal diferencia é que aquí trátase de tomar todas as decisións nun único modelo de optimización.

Trátase dun modelo nun grafo dirixido $G = (N, E)$ no que os nodos representan as posicións iniciais das aeronaves, os puntos de carga de auga, os puntos do incendio e as bases de descanso. O custo asociado ás arestas son o tempo que lle leva a cada aeronave percorrer a distancia entre os nodos. Na Figura 1 represéntase un exemplo de grafo.

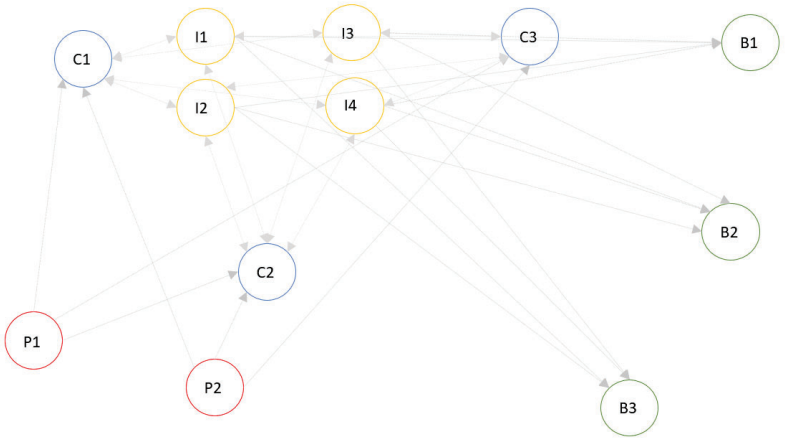


Figura 1: Exemplo de grafo. Os círculos representan os nodos do grafo, nos que s letras I representan os puntos do incendio (puntos de descarga de auga), as letras C os puntos de carga de auga, as letras B representan as bases de descanso das aeronaves e as letras P as posicións iniciais de cada aeronave. As liñas de puntos que unen os nodos son as arestas, que teñen asociado un custo por aeronave que representa o tempo que lle leva a dita aeronave percorrela.

Cada nodo que representa un punto do incendio ten asociada unha eficiencia das descargas de auga establecida previamente polo coordinador do incendio como dato de entrada. Esta eficiencia varía ó longo do tempo, representando deste xeito a evolución do mesmo. Esto represéntase empregando un grafo estendido no tempo, seguindo a idea de Suárez (2016). Na Figura 2 pode verse un exemplo dos nodos do incendio deste grafo estendido.

O coordinador pode establecer un conxunto de preferencias en canto a número de aeronaves que deben traballar nun momento dado, número de litros que deben descargar as aeronaves nun

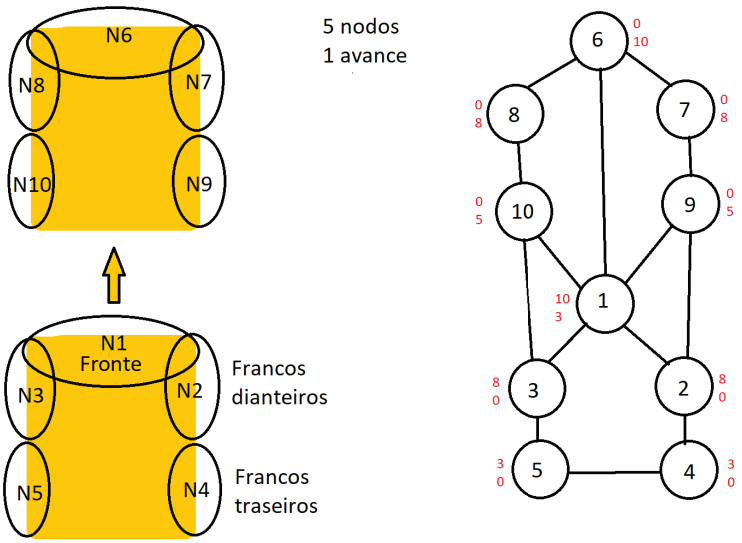


Figura 2: Exemplo dos nodos do incendio estendidos no tempo. A figura da esquerda representa a zonificación do incendio, dividido en cinco partes: a fronte do incendio, dous francos dianteiros e dous francos traseiros. O incendio evoluciona co tempo, e pasa a ocupar outro espacio físico, o que se representa coa parte de arriba, onde antes estaba a fronte do incendio agora está a parte de atrás dos francos traseiros. A representación disto no grafo pode verse na parte dereita da figura. Hai 10 nodos que representan o incendio. No instante inicial, o incendio está situado nos nodos 1-5, mentres que no instante final está situado nos nodos 6-10. Para representar o avance do incendio, o parámetro da eficiencia da descarga de auga asociado a cada nodo varía no tempo. Os números en cor vermella representan esta eficiencia, o número de arriba representa a eficiencia no instante inicial, mentres que o número de abaixo representa a eficiencia no momento final. Así, no instante inicial cando o incendio aínda está nos nodos 1-5, todos os nodos 6-10 teñen unha eficiencia asociada de 0 (se o incendio non se atopa aí non ten sentido descargar auga). Analogamente, no instante final os nodos 2-5 teñen unha eficiencia asociada de 0. O nodo 1 ten unha eficiencia de 3, xa que a parte de atrás do incendio se atopa neste nodo.

punto de incendio ou aeronaves concretas que traballan nun momento concreto. O modelo trata de respectar estas preferencias sempre que sexa posible.

O modelo representa a forma de traballo das aeronaves aproximándose o máis posible á realidade. Entre outras cousas, ten en conta que as aeronaves traballan en grupos determinados polo coordinador (norias), e dentro do mesmo grupo, todas as aeronaves deben cargar e descargar auga nos mesmos puntos. Tamén existe unha limitación da capacidade da auga dos puntos de carga de auga artificiais e unha limitación de capacidade das bases de descanso.

3. HEURÍSTICA

Dada a situación de emerxencia na que se aplicaría este modelo, é preciso que se resolva nun tempo máximo de 10 minutos. Porén, é un modelo complicado, e os tempos de resolución obtidos co solver Gurobi non son aceptables.

Debido a isto, estase a desenvolver unha heurística baseada nunha heurística tipo *simulated annealing* (Gendreau *and* Potvin, 2010; Kirkpatrick *et al.*, 1983). Os movementos implementados pódense dividir en distintos grupos. Por unha banda, hai un conxunto de movementos que consisten en modificar os puntos asignados a un grupo de aeronaves (noria), xa sexa os puntos de carga de auga, de descarga ou as bases de descanso.

Na Figura 3 represéntase o movemento que consiste en modificar as zonas de descarga dunha noria.

Outro movemento consiste en modificar a zona de carga de auga asociada á noria. Neste caso, escóllese ó azar outro punto de carga mantendo o punto de descarga asociado. Na Figura 4 represéntase este tipo de movemento.

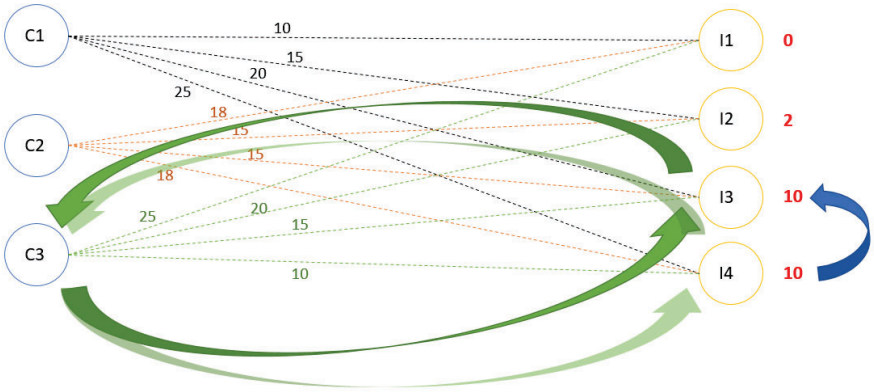


Figura 3: Movementos dos puntos de descarga nunha noria. Os nodos da parte esquerda, etiquetados coa letra C representan os puntos de carga de auga. Os nodos da parte dereita, etiquetados coa letra I, representan os puntos do incendio. Os números sobre as liñas punteadas representan os tempos que unha aeronave tarda en viaxar entre ambos nodos, e os de cor vermella representan a eficiencia de descargar auga no punto do incendio nese instante temporal. Durante a construción da solución inicial, escóllese o punto do incendio que ten unha maior eficiencia e o punto de carga de auga máis próximo. A solución inicial asignaba os nodos C3-I4. Escóllese outro punto aleatorio do incendio e modifícase a zona de descarga asociada á aeronave mantendo a zona de carga. Despois do movemento a aeronave tería asociados os nodos C3-I3.

A Figura 6 representa un dos movementos que se pode facer que consiste en escoller ó azar unha aeronave e modificar o seu intervalo de traballo.

Outros movementos posibles serían ampliar o tempo de traballo dunha aeronave, sempre dentro dos límites establecidos na normativa de voo, ou facer que unha aeronave que está descansando volva ó traballo. Outro conxunto de movementos son os que se encargan de facer que unha aeronave que participa na extinción do incendio non participe, ou o movemento antagónico, que é que unha aeronave que non participe no incendio participe nalgún período de tempo. Na Figura 7 represéntase este movemento.

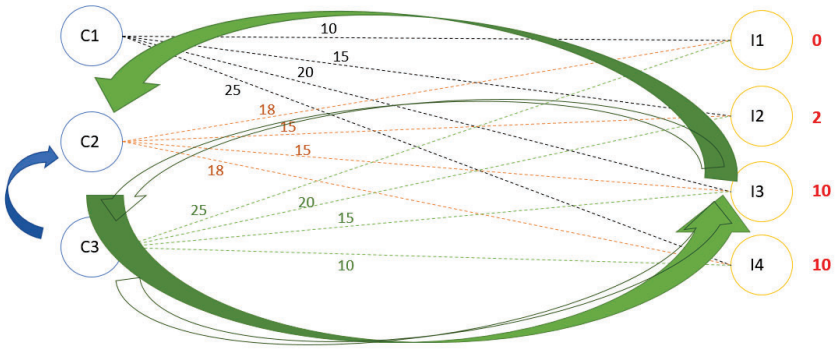


Figura 4: Movementos dos puntos de carga nunha noria. Partindo da situación representada na Figura 3, despois de aplicar o movemento do punto de descarga, escóllese outro punto de carga ó azar. Deste xeito, os novos puntos asociados á aeronave son o punto de carga C2, e o punto do incendio I3.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
n1	a1		C1		I1		B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1
	a2	P2	P2	P2	P2	P2	P2		C1		I1		C3		I4		B1	B1	B1	B1	B1
	a3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3
n2	a4		C4		I3		C4		I3		B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1
	a5	P5	P5	P5	P5	P5	P5	P5	P5	P5	P5		C2		I5		C2		I5		B1

Noria 1:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1	C3	C3	C3	C3	C3	C3	C3	C3	C3	C3
	I1	I1	I1	I1	I1	I1	I1	I1	I1	I1	I4	I4	I4	I4	I4	I4	I4	I4	I4	I4

Noria 2:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C4	C4	C4	C4	C4	C4	C4	C4	C4	C4	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2
	I3	I3	I3	I3	I3	I3	I3	I3	I3	I3	I5	I5	I5	I5	I5	I5	I5	I5	I5	I5

Figura 5: Planificación obtida como solución inicial da heurística. Hai 5 aeronaves, as 3 primeiras (a1, a2 e a3) pertencen á primeira noria (n1) mentres que a4 e a5 pertencen á segunda noria (n2). As columnas representan os intervalos de tempo e as dúas táboas da parte inferior representan os puntos de carga e descarga asociados a cada noria en cada intervalo de tempo.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
n1	a1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1		C3		I4		B1	B1
	a2	P2	P2	P2	P2	P2	P2		C1		I1		C3		I4		B1	B1	B1	B1	B1
	a3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3	P3
n2	a4		C4		I3		C4		I3		B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1
	a5	P5	P5	P5	P5	P5	P5	P5	P5	P5	P5		C2		I5		C2		I5		B1

Noria 1:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1	C3	C3	C3	C3	C3	C3	C3	C3	C3	C3
	I1	I1	I1	I1	I1	I1	I1	I1	I1	I1	I4	I4	I4	I4	I4	I4	I4	I4	I4	I4

Noria 2:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C4	C4	C4	C4	C4	C4	C4	C4	C4	C4	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2
	I3	I3	I3	I3	I3	I3	I3	I3	I3	I3	I5	I5	I5	I5	I5	I5	I5	I5	I5	I5

Figura 6: Movemento aleatorio do intervalo de traballo dunha aeronave. Respecto á Figura 5, a aeronave 1 modifica o seu intervalo de traballo.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
n1	a1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1	P1		C3		I4		B1	B1
	a2	P2	P2	P2	P2	P2	P2		C1		I1		C3		I4		B1	B1	B1	B1	B1
	a3	P3	P3	P3	P3	P3	P3	P3		C1		I4		C3		I4		B2	B2	B2	B2
n2	a4		C4		I3		C4		I3		B1	B1	B1	B1	B1	B1	B1	B1	B1	B1	B1
	a5	P5	P5	P5	P5	P5	P5	P5	P5	P5	P5		C2		I5		C2		I5		B1

Noria 1:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1	C3	C3	C3	C3	C3	C3	C3	C3	C3	C3
	I1	I1	I1	I1	I1	I1	I1	I1	I1	I1	I4	I4	I4	I4	I4	I4	I4	I4	I4	I4

Noria 2:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	C4	C4	C4	C4	C4	C4	C4	C4	C4	C4	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2
	I3	I3	I3	I3	I3	I3	I3	I3	I3	I3	I5	I5	I5	I5	I5	I5	I5	I5	I5	I5

Figura 7: Movemento aleatorio no que unha aeronave que non participaba no incendio empeza a traballar. Respecto á Figura 6, unha aeronave que non estaba participando na extinción do incendio na solución inicial, a aeronave 3, comeza a traballar no incendio nun período de tempo aleatorio.

REFERENCIAS

Consellería do Medio Rural (2023) Plan de Prevención y Defensa contra los Incendios Forestales de Galicia (PLADIGA). Dirección Xeral de Defensa do Monte, Xunta de Galicia.

Dirección General de Aviación Civil (1995) Circular Operativa nº 16-B: Limitaciones de tiempo de vuelo, máximos de actividad aérea y períodos mínimos de descanso para las tripulaciones. Ministerio de Obras Públicas, Transportes y Medio Ambiente, Gobierno de España.

Donovan, G. H., Rideout, D. B. (2003) An Integer Programming Model to Optimize Resource Allocation for Wildfire Containment. Forest Science, 49(2), 331-335.

Gendreau, M., Potvin, J.Y. (2010) Handbook of metaheuristics. New York: Springer.

Granda, B., León, J., Vitoriano, B., Hearne, J. (2023) Decision Support Models and Methodologies for Fire Suppression. Fire, 6, 37.

Kirkpatrick, S., Gelatt Jr, C. D., Vecchi, M. P. (1983) Optimization by simulated annealing. Science, 220(4598), 671–680.

Rodríguez-Veiga, J., Ginzo-Villamayor, M. J., Casas-Méndez, B. (2018) An integer linear programming model to select and temporally allocate resources for fighting forest fires. Forests 9(10), 583.

Rodríguez-Veiga, J., Gómez-Costa, I., Ginzo-Villamayor, M. J., Casas-Méndez, B., Sáiz-Díaz, J. L. (2018) Assignment Problems in Wildfire Suppression: Models for Optimization of Aerial Resource Logistics. Forest Science 64(5), 504–514.

Suárez, D. (2017) A Stochastic Programming Approach for Wildfire Suppression: Pre-positioning and Distribution of Resources Under Uncertainty. Trabajo de grado - Maestría, Universidad de los Andes.

Modelo matemático para la gestión óptima de una planta de regasificación

Ángel M. González Rueda[1], Alfredo Bermúdez de Castro[2], Mohsen Shabani[3] y Christian Álvarez Peláez[3]

[1]Grupo MODESTYA, Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela
[2]CITMAga, Departamento de Matemática Aplicada, Universidad de Santiago de Compostela
[3]CITMAga

RESUMEN

Siguiendo las recomendaciones de la Unión Europea, existe un gran interés en generar mercados de energía más eficientes y limpios. En este contexto surge el interés por el uso del gas natural licuado (GNL) por ser un combustible fósil que emite menos gases de efecto invernadero en comparación con otros tipos de combustibles. Las plantas de regasificación son las encargadas de recibir GNL y regasificarlo convirtiéndolo de nuevo en gas natural, que será enviado a los consumidores finales (tanto industriales como domésticos). En este trabajo, se propone un modelo matemático para la gestión óptima de la operación de una planta de regasificación. El objetivo es minimizar el consumo de energía de la planta de GNL, asociado principalmente a compresores de boil-off gas y bombas, lo cual deriva en una redución de emisiones de gases nocivos. Además, es imprescindible garantizar la seguridad del suministro y cumplir con las restricciones técnicas de los diferentes dispositivos de la planta. El modelo incluye tanto ecuaciones basadas en leyes físico-químicas de los distintos procesos (dadas por ecuaciones algebraicas y diferenciales) como modelos basados en datos experimentales, lo que conduce a la formulación de un problema de optimización polinómico entero mixto.

Palabras y frases clave: planta de regasificación, gestión energética, optimización polinómica entera mixta.

REFERENCIAS

Biegler, L. T., Campbell, S. L., & Mehrmann, V. (Eds.) (2012) Control and optimization with differential-algebraic constraints. Society for Industrial and Applied Mathematics.

Sager, S. (2005) Numerical methods for mixed–integer optimal control problems. Lübeck: Der Andere Verlag. PhD Thesis.

Ye, Z., Mo, X., & Zhao, L. (2021) MINLP model for operational optimization of LNG terminals. Processes, 9(4), 599.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

A UTILIZAÇÃO DO GEOGEBRA COMO METODOLOGIA DE ENSINO E APRENDIZAGEM EM MATEMÁTICA NO ENSINO SUPERIOR
- RESOLUÇÃO DE PROBLEMAS DE PROGRAMAÇÃO LINEAR PELO MÉTODO GRÁFICO -

Carla Martinho[1,2] , Manuel Martins[1]

[1] Instituto Superior de Contabilidade e Administração de Lisboa, Instituto Politécnico de Lisboa
[2] ICPOL - ID&I Unit (PORTUGAL)

RESUMO

Há mais de uma década que as tecnologias de informação e comunicação estão a transformar os processos de ensino e aprendizagem. No entanto, ainda nem todos dominam e integram as novas ferramentas de comunicação e gestão da informação na prática educativa, pois esta exige a aquisição de novas competências profissionais e uma forte aposta na sua inovação. Neste trabalho pretende-se apresentar e analisar como foi realizada a introdução de um software online de acesso livre e gratuito, o GeoGebra, no ensino e aprendizagem numa unidade curricular de matemática em conteúdos de Programação Linear, no 1º ano de uma licenciatura em Gestão do ensino superior politécnico em Portugal.

Palavras e frases chave: ensino superior, GeoGebra, programação linear, resolução gráfica.

1. INTRODUÇÃO

A unidade curricular (UC) de Matemática numa Instituição de Ensino Superior sofreu no ano letivo, 2022/23, uma alteração nos seus conteúdos programáticos, tendo sido introduzido pela primeira vez um ponto relativo à Otimização Linear com duas variáveis. Sabe-se que a otimização assenta essencialmente no estudo de problemas matemáticos em que se pretende maximizar ou minimizar uma função sujeita a um conjunto de restrições, equações ou inequações, com o intuito de encontrar a solução ótima do problema. Sabe-se ainda, que a otimização linear é talvez o problema de otimização mais importante e muito utilizado na área da Investigação Operacional.

Na licenciatura à qual pertence esta UC existem duas UC de investigação operacional, motivo pelo qual a introdução à Programação Linear se realiza, nesta UC, apenas com duas variáveis assentando por isso, a resolução desses problemas, única e exclusivamente, no método gráfico.

A utilização de calculadora deixou, este ano letivo, de ser autorizada, sem que a professora investigadora tenha qualquer responsabilidade nesta decisão. Os alunos deste 1º ano, fizeram todo o seu percurso de ensino e aprendizagem, no básico e secundário, em matemática com recurso à calculadora, pelo que a professora investigadora entende que a sua não utilização os pode desmotivar e pode nalguns casos agravar o "não gosto" pela matemática levando mesmo ao abandono da UC e ao aumento das taxas de insucesso no ensino superior. Acredita-se, por isso, que a utilização do GeoGebra como metodologia de ensino e aprendizagem na introdução à Programação Linear pode ajudar a ultrapassar algumas resistências e ser facilitadora de uma aprendizagem significativa neste domínio.

Assim, este estudo tem como objetivo principal averiguar como alunos do 1º ano do ensino superior resolvem problemas de Programação Linear, com e sem recurso ao GeoGebra. Mais especificamente, pretende-se analisar e compreender como é que os alunos interpretam o enunciado de problemas de Programação Linear, quais as estratégias que adotam, como interagem entre si e com a professora investigadora durante a resolução destes problemas e em que medida como é que utilizam o GeoGebra para resolver os mesmos.

Com a finalidade de conseguir uma articulação entre a investigação didática de Programação Linear e o ensino aprendizagem de Programação Linear no ensino superior, estabeleceram-se atividades de

resolução de exercícios com recurso ao GeoGebra, sendo concretizadas pelos alunos em trabalho colaborativo e com o auxílio da professora investigadora.

Optou-se por isso, por uma metodologia de natureza qualitativa que observa a realização de um estudo de caso particular em sala de aula, investigando o processo ensino e aprendizagem contemporâneo, num contexto real.

Todo o trabalho desenvolvido foi ainda, partilhado pelos alunos num fórum para o efeito, na plataforma Moodle, onde se encontra a página de apoio à UC.

2. INTRODUÇÃO À PROGRAMAÇÃO LINEAR COM RECURSO ÀS TECNOLOGIAS

De acordo com Barros et al. (2010) na sociedade atual existem várias situações em que temos de tomar decisões de planeamento ou de gestão de forma a rentabilizar os recursos disponíveis e minimizar os custos ou consumos, assim é necessário resolver problemas de otimização. Entende-se que para esse tipo de resolução de problemas todas as funções envolvidas (função objetivo e restrições) são lineares e daí temos um problema de PL.

Para Fagundes et al. (2016) a Programação Linear procura encontrar a melhor solução (solução ótima) para problemas que têm os modelos representados por expressões lineares. Esta característica, de linearidade das expressões, torna a PL simples e altamente aplicável.

Ainda de acordo com Fagundes, a representação gráfica de problemas de PL só é possível, quando os problemas apresentam duas variáveis de decisão. A solução ótima de um problema de programação linear pode ser determinada, numericamente, através do Método Simplex.

Barros et al. (2010) complementa que a PL é uma técnica matemática utilizada para resolver problemas de otimização em que há uma relação linear entre as variáveis envolvidas. Em ciências empresariais, a resolução de problemas de programação linear pode ser extremamente importante, pois ajuda a tomar decisões estratégicas com base em informações quantitativas.

De acordo com Joly et al. (2015) a abordagem à Programação Linear (PL) torna-se, por isso, importante para que os alunos adquiriram conhecimentos teóricos e competências práticas sobre o desenvolvimento de modelos de otimização, técnicas de solução e ferramentas baseadas em PL relacionadas com atividades do dia-a-dia e para que possam, ainda, desenvolver competências que permitam tomar decisões devidamente fundamentadas.

Ao trabalhar com conceitos matemáticos relacionados com PL de acordo com Fagundes et al. (2016) tem-se por base uma parte da investigação operacional no seu caráter mais dinâmico, na otimização de determinada situação do problema do cotidiano. Estes, podem ser resolvidos, utilizando ferramentas gráficas, algébricas e computacionais, permitindo, assim, reconhecer a tecnologia como uma importante aliada no processo de ensino e aprendizagem da Investigação Operacional.

Segundo Sprovieri et al. (2021) os problemas de programação linear de duas variáveis, podem ser resolvidos através de algoritmos envolvendo formas de resolução distintas ou combinadas: a resolução gráfica e a resolução algébrica. Este estudo considera que o papel assumido pela representação gráfica é de fundamental importância para o processo de ensino e aprendizagem.

No entanto, é importante notar que a resolução gráfica pode ser limitada em problemas com muitas variáveis e restrições mais complexas. Não sendo este tipo de problemas objeto de análise neste trabalho. Nesses casos, métodos mais eficientes, como o método simplex, são necessários para encontrar a solução ótima de forma mais rápida e precisa. Portanto, é importante que os alunos de ciências empresariais também desenvolvam competências em técnicas de resolução mais avançadas para lidar com problemas complexos do mundo real.

Existem vários recursos educativos digitais utilizados no ensino e aprendizagem de PL que promovem e estimulam o ensino em matemática, criando ambientes de aprendizagem dinâmicos e visuais. A plataforma Moodle, uma das mais populares entre as IES em Portugal, é um software online livre, de apoio ao ensino e aprendizagem, que disponibiliza entre outros, um conjunto de ferramentas de comunicação (fóruns, *chats*), onde os alunos, segundo Salvador et al. (2016), podem apoiar o seu processo de aprendizagem. Em particular, os fóruns são vistos como ferramentas de aprendizagem colaborativa, em que se compartilham experiências, fazem perguntas e se cria um envolvimento na construção de uma comunidade de aprendizagem e prática.

De acordo com Castro, C. (2014), sendo recente o mais intenso apetrechamento informático e tecnológico nas escolas portuguesas com a implementação do Plano Tecnológico da Educação, sabe-se ainda muito pouco sobre o impacto dessa ação e, mais especificamente, nas práticas dos

professores. Para tentar perceber o grau de utilização de recursos educativos digitais (RED) e os fatores que determinam o seu uso pedagógico, efetuámos este estudo exploratório e descritivo.

No âmbito do problema desta investigação, considerámos o desenvolvimento profissional dos professores centrado nas competências digitais. Estas permitem-lhes recorrer à integração das tecnologias de informação e comunicação (TIC) e à utilização de RED, no processo de ensinar e aprender, no sentido de tirar partido da tecnologia presente nas salas de aula portuguesas.

De acordo com Barros et al. (2010) torna-se imprescindível o recurso a ferramentas tecnológicas, não só por constituírem um objeto de motivação no ensino e aprendizagem, como também, pelo facto dos problemas de PL, oriundos de situações reais, nem sempre serem de fácil e rápida resolução, podendo envolver um considerável número de variáveis ou restrições. Refere-se que a sua utilização carece de predisposição para a aprendizagem, permitindo que os alunos resolvam uma maior diversidade de problemas e que os explorem com maior profundidade.

O GeoGebra é um software de matemática dinâmica que combina recursos de geometria, álgebra, cálculo e gráficos num único programa. Segundo Lavicza el at. (2020), foi desenvolvido para auxiliar no ensino e aprendizagem da matemática, permitindo que os seus utilizadores explorem conceitos matemáticos de forma interativa.

Bedada et al. (2022) reforça que no âmbito educacional tanto os educadores como os alunos beneficiam das vantagens da tecnologia, pois os alunos são atraídos por este meio de aprendizagem visualmente divertido e interativo. Acresce que a interação entre a tecnologia e o uso do GeoGebra em sala de aula é um potenciador das aprendizagens em matemática, reforçando o processo de significação do conteúdo.

Para Arbain et al. (2015) a eficácia do GeoGebra está na sua gratuitidade e disponibilidade online, beneficiando o ensino e aprendizagem da matemática em sala de aula, uma vez que diversifica a forma de ensinar.

Acresce que o desafio maior dos professores ao ensinar matemática é explorar problemas complexos que envolvem as teorias e fórmulas de modo a que os alunos aprendam, por isso, o uso do software Geogebra pode aumentar o interesse, confiança e motivação dos alunos pela matemática.

Segundo Sousa el at. (2018), o recurso ao GeoGebra na resolução de problemas de PL, tem sido recorrente e permitiu melhorar as capacidades de entendimento da matemática, sobretudo como uma ferramenta poderosa no auxílio ao ensino e aprendizagem.

Segundo, Lavicza et al. (2020), a integração tecnologia e educação, possibilita garantir uma aprendizagem significativa, permite aos alunos criarem modelos matemáticos das situações de otimização e explorarem as soluções de forma gráfica e interativa. O GeoGebra, configura-se como um software de código aberto e matemático de alta qualidade, um sistema disponível na maioria das plataformas tecnológicas e oferecido gratuito tanto para professores quanto para os alunos.

Molnár (2016) afirma que em 2001, Markus Hohenwarter desenvolveu o sistema GeoGebra, cresceu e foi desenvolvendo novos módulos e novas funções. Atualmente, a versão (versão 5.0) oferece aos utilizadores a janela de geometria 3D, que permite trabalhar com sólidos no sistema de coordenada tridimensionais. O GeoGebra pode aceder-se através de uma plataforma móvel permitindo, por isso, a sua utilização em tablets, computadores e telefones móveis.

Ensinar matemática sempre apresentou muitos desafios, tendo estes crescido nos últimos anos com o avanço das tecnologias e o menor interesse dos alunos no ensino e aprendizagem da matemática, apoiado em políticas da educação, nem sempre promotoras do mesmo, junto das novas gerações, ainda mais quando confrontados com matérias mais avançadas e abstratas como no ensino superior. Por isso, são vários os autores, que há mais de uma década desenvolvem investigações no ensino da matemática com recurso ao GeoGebra, (Molnar e Lukay, 2015), (Kriek e Stols, 2011) , (Ainley et al., 2010).

3. METODOLOGIA E ANÁLISE DOS RESULTADOS

A Introdução à Programação Linear passou a fazer parte do programa da unidade curricular (UC) de matemática II, do 1º ano, das licenciaturas de Gestão, Finanças e Contabilidade, do Instituto Superior de Contabilidade e Administração de Lisboa, do Instituto Politécnico de Lisboa, instituição do ensino superior politécnico público português. O seu conteúdo corresponde a um terço da duração, em horas, da mesma UC. Mais concretamente, no tema 2 do conteúdo programático, relativo à Programação Linear: Definições e conceitos básicos. Formulação de problemas em Programação Linear: exemplos. Resolução gráfica de problemas com duas variáveis. Apresentação e resolução de alguns problemas. Dualidade.

Análise de sensibilidade: abordagem gráfica. A professora investigadora desenvolveu o seu estudo no processo de ensino e aprendizagem, em contexto de sala de aula, sobre este tópico da matéria.

Para tal, a presente investigação realizou-se com os alunos de duas turmas do 1º ano, regime diurno, da licenciatura em Gestão do Instituto Superior de Contabilidade e Administração de Lisboa (ISCAL), tendo por objetivo analisar como estes resolvem problemas de Programação Linear pelo método gráfico, com e sem recurso ao GeoGebra e em que medida a utilização deste software contribui para a melhoria do ensino e aprendizagem, bem como, motivar os alunos e ser facilitador de uma aprendizagem significativa neste domínio.

Como já referido, este ano letivo, nas referidas licenciaturas a utilização de calculadora deixou de ser autorizada nesta UC, sem que a professora investigadora tenha qualquer responsabilidade nesta decisão. Assim, foram analisadas as seguintes questões:

a. Como os alunos interpretam o enunciado dos problemas de PL?
b. Quais as estratégias utilizadas na resolução dos problemas de PL sem software e sem calculadora?
c. Como interagem entre si e com a professora investigadora durante a resolução destes problemas?
d. Em que medida a resolução de problemas de PL com recurso ao GeoGebra e ao trabalho colaborativo impactou o processo de ensino e aprendizagem e a motivação dos alunos?

O método utilizado para a realização deste estudo foi o de investigação-ação colaborativa, caracterizado por procurar simultaneamente investigar e resolver um problema. Este método é recorrente nas investigações em sala de aula, realizado por professores investigadores, por serem simultaneamente académicos, com experiência em trabalhos de investigação, e professores, com conhecimento e domínio da prática pedagógica. (Wright, 2021).

Os professores investigadores utilizam esta forma sistemática de investigação, por dar prioridade à reflexão e fazer a ponte entre a teoria e a prática. Pela sua natureza, muitas vezes, é também designado por ciclo de ação ou ciclo de investigação, que conforme se apresenta na figura pode ser descrito em quatro etapas:

Figura 1. Ciclo de ação ou ciclo de investigação
Fonte: Adaptado de George, 2023

Na fase de planeamento definem-se os objetivos, os problemas são identificados e analisados, e são elaboradas estratégias e planos de ação para abordar as questões identificadas. Essa etapa é essencial para estabelecer uma base sólida para a implementação da ação.

Na fase da ação, implementam-se as estratégias e os planos estabelecidos na etapa de anterior. Nesta etapa, as mudanças e intervenções planeadas são realizadas no contexto em questão.

A fase de análise ocorre após a implementação da ação, sendo essencial observar e monitorizar cuidadosamente os resultados e efeitos da intervenção. Os dados e informações relevantes são recolhidos, conduzindo aos resultados que permitirão analisar e avaliar o impacto das ações concretizadas.

A última fase, fase da conclusão, inclui a reflexão envolvendo uma análise crítica do processo, dos resultados alcançados e do que pode ser melhorado. Com base na avaliação, novos *insights* são obtidos, e o ciclo recomeça com novas etapas de planeamento, ação e assim por diante.

Segundo Leat et al. (2014) os professores que se envolvem totalmente com a investigação têm mais hipóteses de gerar novos *insights* e perspetivas, efetuando desta forma fortes mudanças na sua prática, tornando-se simultaneamente mais críticos em relação às políticas e práticas existentes.

Assim, a investigação-ação colaborativa é uma abordagem que promove a participação ativa dos alunos no processo de ensino e aprendizagem e os incentiva a refletir sobre o conhecimento adquirido. Esta foi a estratégia utilizada para esta experiência: Enquanto professora investigadora, tentar melhor compreender a dificuldade dos alunos em relação a matemática e ao seu processo de ensino e aprendizagem.

Atividades desenvolvidas

Todo o processo de investigação, foi realizado, em contexto de sala de aula com os alunos presentes em cada uma das duas turmas. Estas aulas foram assistidas por uma colega da professora investigadora que ia registando o processo fotograficamente e tirando algumas notas descritivas da ação. Para que tal acontecesse, foi solicitado aos alunos presentes em aula, autorização prévia, para esse registo.

As aulas decorreram entre a 6ª e a 9ª semana letiva do 2º semestre de 2022/23, nas datas e com os conteúdos, de acordo com a planificação apresentada no tabela 1:

Tabela 1. Planificação das aulas de Programação Linear

Semanas	Conteúdos
6ª (27 mar / 01 abr)	Programação linear. Definições e conceitos básicos. Formulação de problemas em programação linear. Exemplos. Resolução gráfica de problemas com duas variáveis. Introdução ao GeoGebra.[1]
(2 abr/10 abr)	Férias da Páscoa
7ª (11 abr / 15 abr) Páscoa – 2ª feira	Dualidade. Construção do problema dual. Teorema fundamental da dualidade. Teorema dos desvios complementares (sem demonstração). Interpretação económica. Resolução de exercícios.
8ª (17 abr / 22 abr)	Pós-otimização e análise de sensibilidade. Abordagens gráfica: Alteração nos coeficientes da função objetivo; Alteração nos termos independentes das restrições; Introdução de uma nova restrição. Resolução de exercícios.
9ª (24 abr / 29 abr) Feriado - 3ª feira	Resolução de exercícios. 1º Teste

Cada semestre letivo é composto por 15 semanas, tendo a UC de Matemática II uma carga horária de 4,5 horas semanais. As aulas realizavam-se às 3ª e 4ª feiras, em blocos de 100 minutos. Cada uma das turmas tinha por isso, três blocos de 100 minutos. A turma A tinha dois blocos à 3ª feira e um bloco à 4ª feira, e a turma B tinha um bloco à 3ª feira e dois blocos à 4ª feira. Para o tópico 2 do programa de Matemática II, relativo à Programação Linear, ficaram prevista as 4 semanas correspondentes à 6ª, 7ª, 8ª e 9ª semana do semestre em curso.

[1] A Introdução ao GeoGebra, não estava prevista na planificação inicial da UC.

Para que os alunos pudessem realizar as tarefas planeadas pela professora investigadora com recurso ao GeoGebra, esta teve de incluir na 6ª semana, de 27 de março a 2 de abril, um tópico de introdução ao GeoGebra, inicialmente não previsto.

Além, disso a professora investigadora, convidou os alunos a participarem numa sessão específica para exploração do GeoGebra, na sala de computadores. Esta sessão realizou-se na 8ª semana, a 19 de abril, fora do horário letivo das aulas, das 14 às 16 horas.

Todas as aulas planificadas, combinaram duas componentes com diferentes tempos a elas destinadas: aula teórica com cariz mais expositivo, aula prática de trabalho colaborativo entre os alunos com cariz mais voltado para a aplicação dos conceitos teóricos em situações concretas, através da resolução de exercícios práticos e problemas do mundo real. Neste formato de aula, o foco está na prática, na resolução de problemas e no desenvolvimento das competências dos alunos em aplicar os conhecimentos matemáticos apreendidos.

Uma aula prática de trabalho colaborativo entre os alunos é uma abordagem que enfatiza a cooperação, a participação ativa e a troca de conhecimentos entre os estudantes. Neste formato de aula, a professora investigadora assumiu um papel de facilitador, incentivando os alunos a trabalharem juntos para resolver problemas e alcançar objetivos comuns, conforme descrito por Jagtap (2016).

Esta estratégia foi pensada especificamente, no contexto apresentado, por serem turmas grandes, com mais de 25 alunos por aula, e por existir um caderno com muitos exercícios propostos.

Além das aulas planificadas, a professora investigadora contou com o apoio da plataforma Moodle, para onde os alunos estendiam os trabalhos e desafios começados dentro da sala de aula.

Descrição das aulas ministradas

6ª semana - aula de 28 de março de 2023 – 1ª aula planificada para PL (duração 100 min)

Por se tratar de uma aula de introdução ao tema, tendo por isso, um carácter mais expositivo na primeira parte da aula a professora investigadora optou por juntar as duas turmas num auditório e fez uma apresentação alargada sobre Programação Linear, incidindo nos problemas com duas variáveis e na importância desta no contexto do curso de Gestão.

Figura 2. - 1ª aula de PL, 28 de março de 2023, no auditório com as duas turmas

Num primeiro momento, foi realizada a introdução aos primeiros conceitos da Otimização linear. Os alunos ouviram e observaram toda a explicação baseada em slides e também a apresentação de alguns livros físicos e outros disponíveis online, de referencia ao estudo do conteúdo temático planificado.

A par desta exposição, a professora investigadora ia questionando os alunos sobre o entendimento e compreensão dos conceitos apresentados e ia solicitando a sua participação para encontrar exemplos da vida real, nomeadamente quando introduzir a formalização de problemas.

O material produzido pela professora investigadora, foi efetuado tendo por base os conceitos-chaves e as definições em relação à programação linear com duas variáveis, de acordo com a planificação anteriormente apresentada na tabela 1 e com a ficha da unidade curricular (FUC): Seguindo, por isso, a seguinte estrutura:

Figura 3. Índice de slides de Programação Linear da UC de Matemática II

A professora investigadora referiu nesta introdução a frase de Pólya "*A Matemática não é um desporto para espectadores: não a podemos fazer, nem aprender, sem uma participação ativa."* Apresentou aos alunos a etapa de um estudo de PL, exemplificando com um problema concreto e despertando-os para a necessidade da interpretação dos enunciados por forma a que conduzam à formulação correta dos mesmos, utilizando para o efeito o seguinte exemplo:

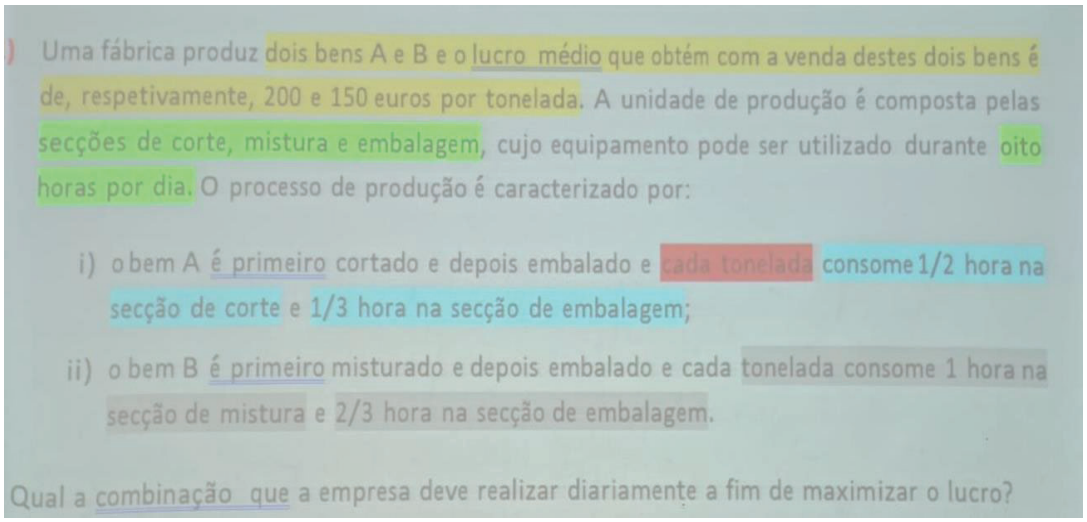


Figura 4. Exemplo de interpretação de enunciado

No final desta aula, a professora investigadora sugeriu aos alunos que discutissem entre si e tentassem formular dois problemas do caderno de exercícios de suporte a esta temática.

6ª semana - aula de 29 de março de 2023 – 2ª aula planificada para PL (duração 200 min)

A aula iniciou-se com a discussão dos problemas que alguns dos alunos tinham tentado formular, tendo a professora investigadora realizado , passo a passo, a formulação detalhadamente para toda. A aula prosseguiu com a constituição de grupos de trabalho, para que dessem continuidade à interpretação dos enunciados e efetivassem novas formulações de forma colaborativa.

Figura 5. Grupos de trabalho colaborativo

A professora investigadora foi circulando pela sala interagindo e respondendo às questões colocadas pelos alunos, sublinhando nos seus próprios cadernos as frases conducentes à identificação de cada restrição ou à função objetivo.

Utilizando os exercícios formulados pelos alunos, a professora investigadora deu início à resolução de problemas pelo método gráfico, aproveitando para fazer uma revisão do traçado de retas e determinação do semiplano associado a cada inequação. Foram em seguida, explanados mais alguns conceitos teóricos, revisitados outros, e explicou-se a resolução dos problemas pelo método gráfico, passo a passo, como se exemplifica na figura seguinte:

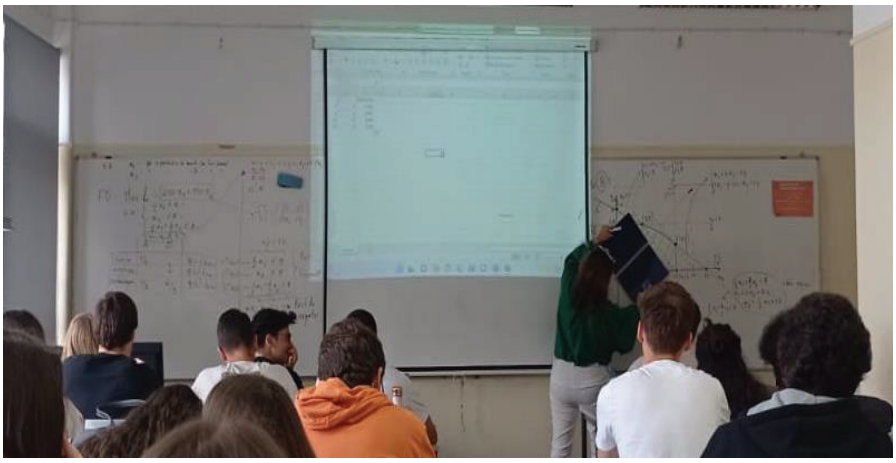


Figura 6. Exemplo de resolução do problema pelo método gráfico

Para o traçado de curvas de nível, retas paralelas, a professora investigadora, sem régua nem esquadro adaptado ao quadro, procurou nos materiais disponíveis em cima das mesas dos alunos algum objeto que a auxiliasse nessa tarefa: recorrendo a um caderno de espiral A4, que quando aberto permitia visualizar o vetor gradiente em relação ao qual era necessário traçar as retas perpendiculares. Este momento, causou algum burburinho entre os alunos acabando por ser um momento divertido em que os alunos perceberam a importância do material, pois sem ele, a representação gráfica realizada pela professora investigadora no quadro ficaria menos precisa e seria mais difícil determinar graficamente a solução ótima.

A aula prosseguiu com os mesmos grupos de trabalho, dando continuidade à resolução de problemas pelo método gráfico, com a ajuda da professora investigadora.

Figura 7. Trabalho colaborativo de resolução de PPL

Depois de vários exercícios resolvidos pelos alunos, a professora investigadora, aproveitando algumas questões colocadas sobre a solução ótima do problema, introduziu o GeoGebra por forma a mais rapidamente poderem visualizar e encontrar as respostas a estas questões, permitindo assim, melhor compreendendo os conceitos associados. Aproveitou o exercício inicialmente resolvido no quadro, para fazer nova resolução, agora, com recurso ao GeoGebra, conforme se apresenta na figura seguinte:

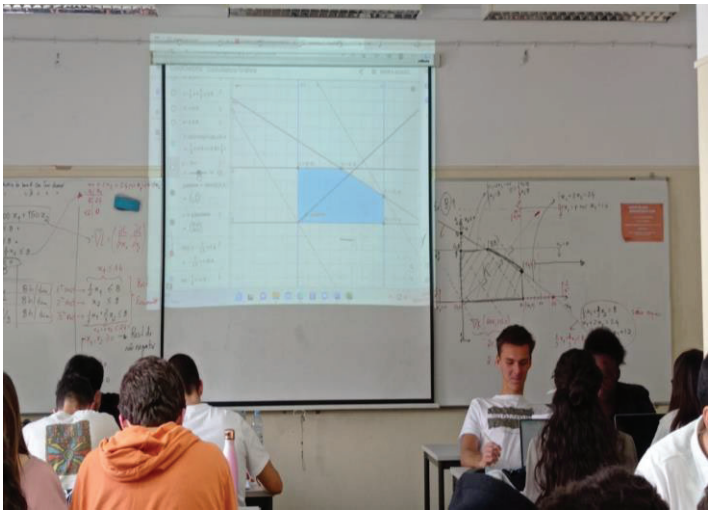


Figura 8. Iniciação ao GeoGebra

Alguns grupos de alunos utilizaram os seus computadores portáteis para ir acompanhando e desde logo explorando o software apresentado. Como se apresenta na figura seguinte:

Figura 9. Resolução de problema de PL pelo método gráfico com recurso ao GeoGebra

A professora investigadora abriu um fórum, para as duas turmas, destinado a colocarem questões e a partilharem o trabalho realizado colaborativamente em sala de aula, com a supervisão da professora investigadora, que previamente orientou os alunos na digitalização de documentos utilizando uma app apropriada para o telemóvel. A adesão foi imediata, como se pode verificar pela figura:

Figura 10. Partilha de PPL no fórum da plataforma Moodle, 30 de março

7ª semana - aula de 11 de abril de 2023 – 3ª aula planificada para PL (duração 200 min)

A segunda semana de aulas relativo ao tema, realizou-se depois das férias da Páscoa, pelo que a professora investigadora iniciou a 3ª aula chamando à atenção para algumas imprecisões detetadas na resolução dos PPL partilhados no Moodle, enfatizando a importância da precisão na representação gráfica, para encontrar a solução ótima. Explicou, mais uma vez, passo a passo a metodologia para determinar o conjunto de soluções admissíveis.

Figura 11. Partilha de exercícios resolvidos no fórum da plataforma Moodle

Dando continuidade ao trabalho colaborativo entre alunos para a resolução de problemas de PL pelo método gráfico. Pediu a colaboração de um aluno para resolver um problema no quadro, explicitando oralmente o seu raciocínio:

Figura 12. Colaboração de um aluno para resolver um PPL no quadro

Tendo percecionado, que os alunos se perdiam com alguma frequência na determinação do conjunto de soluções admissíveis (CSA), elaborou um esquema que orienta-se a determinação do semiplano correspondente a cada restrição e conduzisse à representação gráfica exigida e conducente ao CSA, conforme se apresenta nas figuras seguintes:

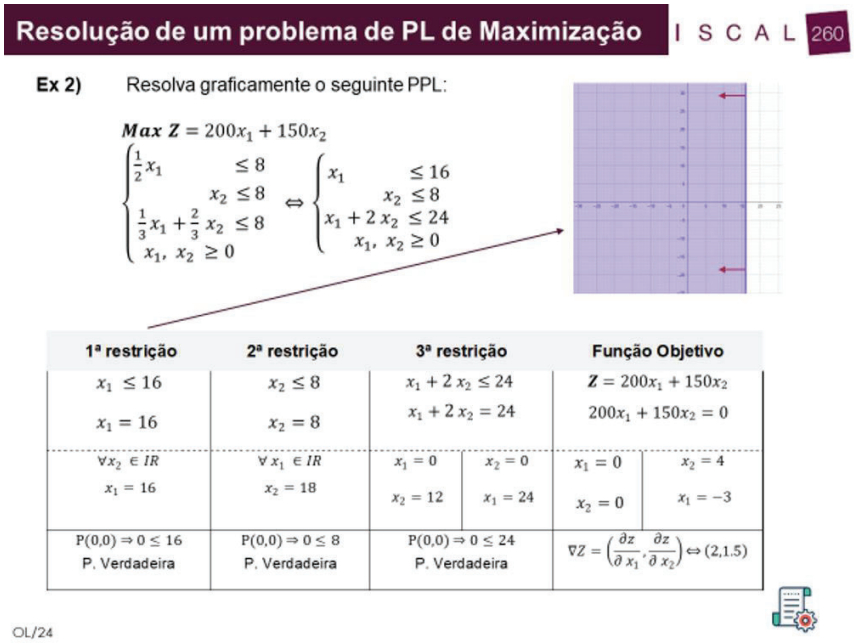


Figura 13. Esquema orientador da representação gráfica na resolução de um PPL

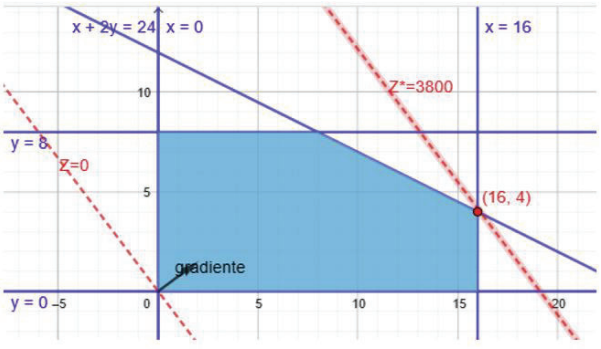


Figura 14. Determinação do conjunto de soluções admissíveis (CSA)

Nesta sessão foi ainda solicitado aos alunos presentes que realizassem uma ficha, com conceitos básicos de representação gráfica.

- **7ª semana - aula de 12 de abril de 2023 – 4ª aula planificada para PL**
  (duração 100 min)

Nesta aula a professora investigadora iniciou o estudo da dualidade e prosseguiu com o trabalho colaborativo de resolução de PPL, quer em sala de aula quer com recurso ao fórum da plataforma Moodle. Nesta altura, a maioria dos alunos a frequentar as aulas já recorria ao GeoGebra para verificar as soluções encontradas na resolução à mão. No entanto havia, ainda, alunos com muitas dificuldades nas representações gráficas, dando conta à professora investigadora que a maior ajuda era o trabalho colaborativo, que continuava a perdurar, conforme se ilustra na figura seguinte:

Figura 15. Partilha de PPL no fórum da plataforma Moodle, 12 de abril

8ª semana - aula de 18 de abril de 2023 – 5ª aula planificada para PL (duração 300 min)

A professora investigadora iniciou a aula com a utilização do quadro interativo, para recuperar uma resolução realizada à mão e gravada no seu computador portátil, e explicitou o conceito de correspondência entre os problemas primal e o dual, através de um problema já resolvido:

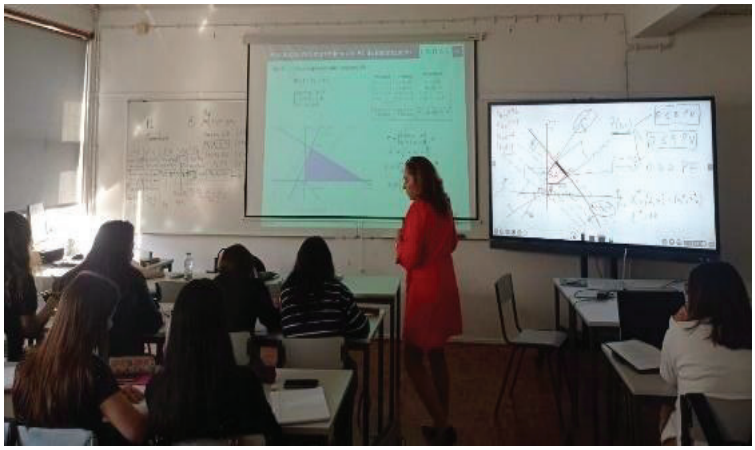


Figura 16. Determinação da solução ótima dos problemas primal e do dual

Os alunos resolveram colaborativamente alguns PPL em que era possível a resolução do primal e do dual pelo método gráfico e posteriormente estabeleceram com a professora investigadora a correspondente relação entre as restrições e variáveis entre ambos. Permitindo, desta forma, verificar que essa relação possibilita, inclusivamente, encontrar a solução ótima, quando existe, de um dos problemas, com base na solução ótima do outro. Construíram o problema dual e a professora investigadora enunciou o teorema fundamental da dualidade e o teorema dos desvios complementares, tendo aproveitado as resoluções efetuadas pelos próprios alunos, supervisionados pela professora investigadora.

Figura 17. Trabalho colaborativo supervisionado pela professora investigadora

8ª semana - aula de 19 de abril de 2023 – 6ª aula planificada para PL (duração 100 min)

A professora investigadora inicia a aula com um problema proposto pelos alunos. No decurso da resolução vai relembrando conceitos fundamentais da PL e retomando as principais dúvidas colocadas pelos estudantes.

A aula prossegue com alguns alunos a preferirem trabalhar sozinhos e outros a trabalhar colaborativamente, realizando análise de sensibilidade com recurso ao GeoGebra e a discutirem a solução encontrada.

Figura 18. Trabalho autónomo/colaborativo supervisionado pela professora investigadora

8ª semana - aula de 19 de abril de 2023 – aula extra na sala de computadores (duração 100 min)

Existiam alunos que, ainda, não estavam familiarizados com o GeoGebra, a professora investigadora tinha planeado uma aula extra para facilitar a sua utilização, tendo por isso orientado a abertura de conta, para que pudessem guardar o trabalho realizado nesta sessão e retomá-lo mais tarde.

Mostrou como criar gráficos de funções lineares, representar as restrições e encontrar a interseção das retas para encontrar o ponto ótimo.

Aproveitou um problema de programação linear resolvido à mão, numa das aulas anteriores, pelo método gráfico para o resolver com recurso ao GeoGebra.

$$\begin{aligned} \textbf{Max } Z = & \; x_1 - 2x_2 \\ s.a \quad & x_1 + 2x_2 \geq 2 \\ & x_1 \leq 2x_2 \\ & x_1 + x_2 \leq 5 \\ & 2x_2 \leq 7 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Mostrou, passo a passo, como inserir no software:

- a função objetivo;
- as restrições;
- o vetor gradiente e as curvas de nível.

Por último, como determinar e interpretar o ponto de interseção das retas como a solução do problema.

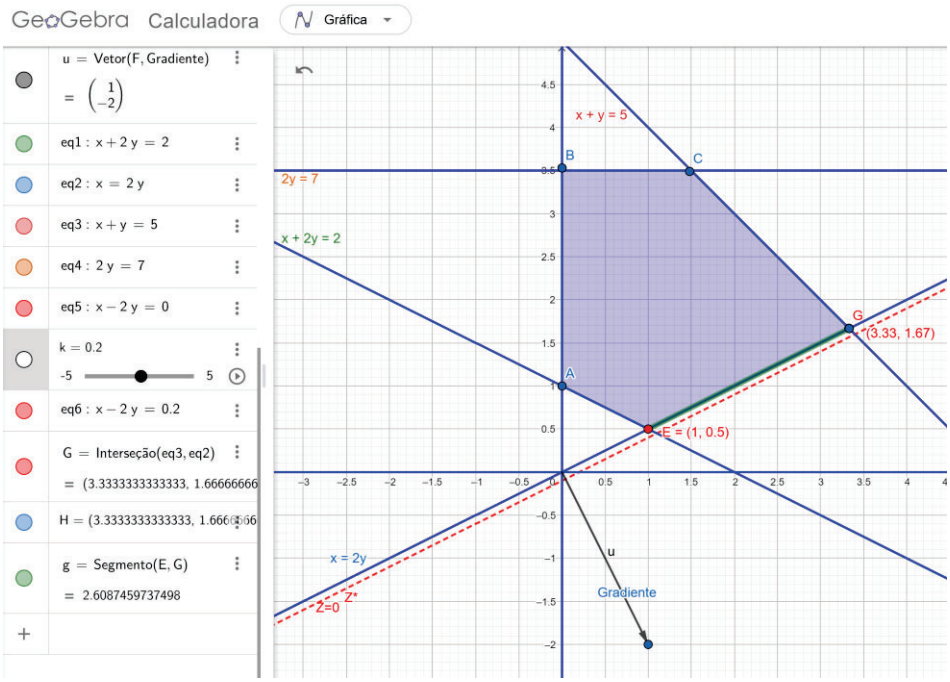


Figura 19. Resolução, passo a passo, de um PPL com recurso ao GeoGebra

Os alunos participaram ativamente:

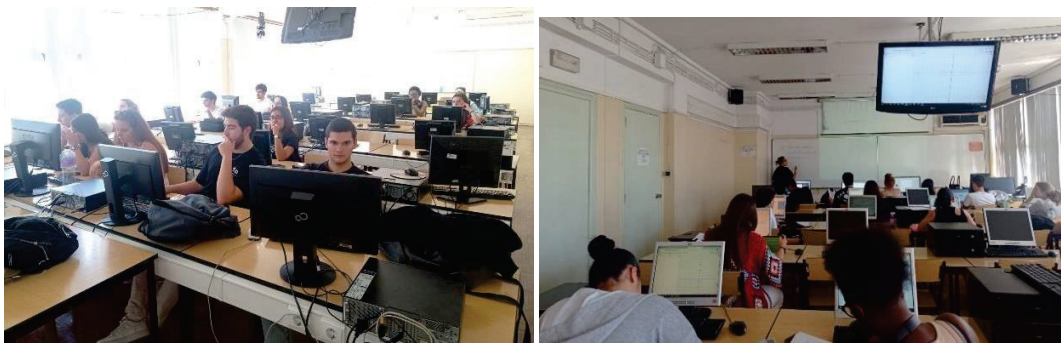


Figura 20. Aula na sala dos computadores

A professora investigadora, foi orientando o trabalho dos alunos:

Figura 21. Trabalho supervisionado, GeoGebra

Nesta sessão foi ainda solicitado aos alunos presentes que realizassem a ficha, já 3ª aula, com conceitos básicos de representação gráfica.

9ª semana - aula de 26 de abril de 2023 – 7ª aula planificada para PL (duração 100 min)

Dia 25 de abril é dia feriado, em Portugal, pelo que esta foi a última aula antes do 1º momento de avaliação contínua. Atendo à realização das 1.ª Jornadas Pedagógicas do ISCAL, as duas turmas tiveram aulas juntas tendo sido efetuadas revisões e esclarecimento de dúvidas.

A professora investigadora, utilizou um problema e foi solicitando aos alunos para elaborarem as perguntas de forma a abarcar toda a matéria alvo de avaliação. Tratou com especial atenção os conceitos de dualidade forte e fraca, reconhecendo as principais diferenças, por haver vários alunos a colocarem questões sobre este assunto. Foi essencialmente uma aula de exposição mas muito participada.

Figura 22. Aula de revisões e esclarecimento de dúvidas

O Moodle continuou em todas as aulas, e para além delas, a ser um apoio ao ensino e aprendizagem de PPL, tendo o fórum criado para o efeito 37 interações, a grande maioria, com partilha de resolução de problemas.

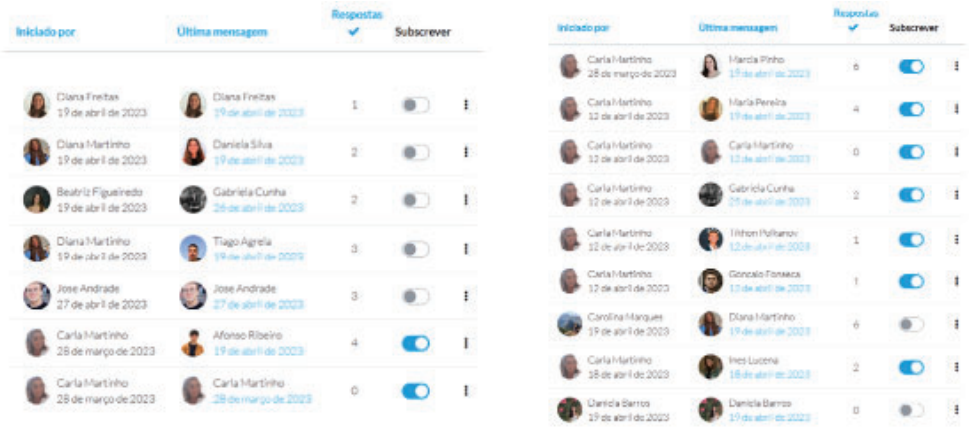


Figura 23. Fórum de apoio à resolução de PPL

Apresentação e análise de resultados

Três turmas do regime diurno, num total de 144 alunos, da unidade curricular de matemática, foram as consideradas para este estudo, todas com níveis de aptidão e históricos semelhantes. Duas turmas (Grupo A e Grupo B) foram designadas como grupos experimentais e utilizaram o GeoGebra e os fórun do Moodle para partilhar e discutir trabalhos colaborativos realizados em sala de aula sob a supervisão da professora. A outra turma (Grupo C) não utilizou esta abordagem, servindo como grupo de controlo.

A avaliação desta UC está definida por dois métodos: i. Avaliação Contínua: dois testes, o primeiro com uma ponderação de 60% e o segundo de 40%; ii. ou exame final de 100%. Sendo que a avaliação contínua exige a pontuação mínima de sete valores no primeiro teste.

Os testes de avaliação contínua, foram realizados, para as três turmas do regime diurno, em simultâneo, sem que a professora investigadora conhece-se o enunciado das provas, em qualquer um dos momentos de avaliação.

O 1º teste, realizado a 29/4/2023, foi constituído por dois grupos, sendo o 2º grupo relativo ao tema de PL, com quatro alíneas, cujos resultados por turma e por alínea se apresentam na tabela seguinte:

Tabela 2. Classificações médias, por alínea, grupo de PL (1º teste)

Turma	2a) (20)	2b) (12)	2c) (13)	2d) (20)	Média Grupo PL
TPGDA	15,10	1,90	12,14	3,95	55%
TPGDB	14,41	2,53	9,31	1,47	46%
TPGDC	9,93	2,34	6,14	3,28	35%

Os resultados médios, por turma, para cada alínea indicam maiores médias em todas alíneas para as turmas sujeitas à metodologia proposta, apresentando, também para ambas, uma classificação média global, neste grupo, superior à da turma C.

Calculou-se e apresenta-se na tabela abaixo a taxa de permanência em avaliação contínua. Este indicador é duplamente importante, primeiramente na prevenção do abandono, muito presente no 1º ano do ensino superior, seguido da não abstenção às aulas. Justificando-se este, pelo facto de, um aluno que tire uma nota inferior a sete fica afastado da avaliação contínua, de acordo com as regras de avaliação em vigor nesta UC, o que aumenta a probabilidade de deixar de frequentar as aulas e consequentemente de não ser aprovado em nenhuma época de avaliação.

Tabela 3. Taxa de permanência em avaliação contínua

Turma	Inscritos	Presenças 1º teste	Presenças 2º teste	Taxa de Perm.
TPGA	45	21	17	81%
TPGB	46	32	25	78%
TPGC	53	29	17	59%

Aqui também ficou evidente que os alunos da turma A e B têm taxas de permanência superiores aos alunos da turma C. Pode por isso, afirmar-se que 41% dos alunos da turma C obteve notas inferiores a 7 valores, *versus* os 19% e 22% das turmas A e B, respetivamente.

Analisou-se ainda os resultados obtidos pelas três turmas findas todas as épocas de avaliação e os resultados são os que se apresentam na tabela seguinte:

Tabela 4. Taxa de êxito e taxa de sucesso

Turma	Inscritos	Presenças 1º teste	Aprovados	Tx Suc	Tx Ex
TPGA	45	21	19	90%	42%
TPGB	46	32	29	91%	63%
TPGC	53	29	17	59%	32%

Considerando a taxa de êxito definida por

$$Tx\ Ex = \frac{nº\ de\ aluno\ aprovados}{nº\ de\ alunos\ inscritos}$$

E taxa de sucesso definida por

$$Tx\ Suc = \frac{nº\ de\ aluno\ aprovados}{nº\ de\ alunos\ presentes\ em\ aula}$$

Pode constatar-se, pela análise da tabela 4, que as turmas que fizeram parte desta investigação, obtiveram maiores taxas de sucesso. Em ambas as turmas, o insucesso foi inferior ou igual a 10% face aos alunos que estiveram presentes em aula. Note-se que a metodologia de ensino e aprendizagem, incluiu para cada ponto da matéria, recursos educativos digitais e apoio nas tecnologias de informação e comunicação adaptados às mesmas.

REFERÊNCIAS

Ainley, J., Eyeleigh, F., Freeman, Ch., O Malley, K. (2010) TIC no ensino de ciências e matemática no ano 8 na Austrália: relatório da segunda pesquisa internacional de estudo de tecnologia em educação (SITES) da IEA, Monografia de Pesquisa ACER, nº 64.

Arbain, N., e Shukor, N. A. (2015) The effects of GeoGebra on students achievement. Procedia-Social and Behavioral Sciences, 172, 208-214. doi: 10.1016/j.sbspro.2015.01.356

Bedada, T. B., e Machaba, M. F.,(2022) The Effect of GeoGebra on Students' Abilities to Study Calculus. Education Research International, 2022, Article ID 4400024, 14 pages. https://doi.org/10.1155/2022/4400024

Barros, P.M., Pereira, A.I., Teixeira, A.P., (2010) À descoberta de software para explorar a programação linear no Ensino Secundário. In ProfMat' 2010, 1-10, Aveiro.

Castro, C. (2014) A utilização de recursos digitais no processo de ensinar e aprender, Prática dos professores e perspetivas dos especialistas. Tese de Doutoramento, Universidade Católica Portuguesa. 414 pp.

Fagundes, H., Pereira, R., Silva, C., Almeida, E., (2016) Programação Linear: Uma proposta de Ensino e Aprendizagem. III CONEDU - Congresso Nacional de Educação.

George, T. (2023) What is action Research? | Definition & Examples. Scribbr. https://www.scribbr.com/methodology/action-research/

Jagtap, P. (2016) Teachers role as facilitator in learning. Scholarly Research Journal, 3(17): 3903-3905.

Joly, M., Rocha, R., Sousa, L., Takahashi M., Mendonça P., Moraes L. e QuelhasS A., (2015) The strategic importance of teaching Operations Research for achieving high performance in the petroleum refining business, Education for Chemical Engineers, 10: 1-19, https://doi.org/10.1016/j.ece.2014.11.001

Kriek, J. e Stols, G. (2011) Por que nem todos os professores de matemática usam software de geometria dinâmica em suas salas de aula? No Australasian Journal of Educational Technology, 27 (1), 137-151. DOI: https://doi.org/10.14742/ajet.988

Lavicza, Z., Prodromou, T., Fenyvesi, K., Hohenwarter, M., Juhos, I., e Koren, B. (2020) Integrating STEM-related Technologies into Mathematics Education at Large Scale. International Journal for Technology in Mathematics Education, 27(1): 3-12. https://doi.org/10.1564/tme_v27.1.01

Leat, D., Lofthouse, R., e Reid, A. (2014) Teachers' views: Perspectives on research engagement. London: Research and Teacher Education: The BERA-RSA Inquiry.

Molnár, P. (2016) Solving a linear optimization word problems by using GeoGebra. ICTE Journal, 5(2): 16-28. DOI: 10.1515/ijicte-2016-0006

Pólya, G. (2004) How to Solve It, A New Aspect of Mathematical Method, Expanded Princeton Science Library Edition, USA.

Salvador, T., Pedrosa, S., Messias, I., Mendes, A.J., e Carvalho E Silva, J., (2016) ReM@t – a project to help students to improve mathematical skills. In International Conference on e-Learning'16.

Sousa, R., Furtado, C. J. G. e Horta, J. C. L. (2018) Resolução gráfica de um problema de programação linear utilizando a folha gráfica 3D do GeoGebra. Revista do Instituto GeoGebra de São Paulo, 7 (2): 45-64, ISSN 2237-9657.

Sprovieri, P. F., e Comelli, C. F., (2021) Deterministic algorithm: a basis for the implementation of educational applications that assist the teaching-learning process in solving two-variable linear programming problems. Revista Interface Tecnológica, 18(2): 290–303, 2021. https://doi.org/10.31510/infa.v18i2.1303

Wright, P. (2021). Transforming mathematics classroom practice through participatory action research. J Math Teacher Educ 24: 155–177. https://doi.org/10.1007/s10857-019-09452-1

Project Funded by IDICA IPL/2022/MOOCS4ALL_ISCAL

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

A RETROSPECTIVE ANALYSIS OF ALCOHOL-RELATED EMERGENCY CALLS TO THE AMBULANCE SERVICE IN GALICIA

Mª José Ginzo Villamayor[1], Paula Saavedra Nieves[1], Dominic Royé[2] and Francisco Caamaño Isorna[3]

[1]Department of Statistics, Mathematical Analysis and Optimization (USC) and Galician Centre for Mathematical Research and Technology (CITMAga)

[2]Climate Research Foundation (FIC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)

[3]Department of Public Health (USC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)

ABSTRACT

This work will be focused on the introduction of statistical methods for data processing and modeling in society, specifically, on alcohol consumption and abuse in Galicia. Dataset is available from a retrospective cohort study based on the telephone calls to the Galicia-061 Public Health Emergency Foundation after alcohol consumption from 1 January 2007 to 4 February 2018. Bayesian hierarchical models and nonparametric level set estimation techniques will applied.

The main objective is modeling spatial and spatio-temporal patterns of emergency calls to the department ethyl poisoning in this region. By fixing administrative areas, for example, municipalities, spatial and spatio-temporal methods for counting data can be considered in this setting. This approach allows to allow to study the evolution of callings patterns. Specifically, hierarchical modeling, through Besag York Mollìe (BYM) method will be used to meet this goal (see Besag *et al.* (1991) and Rue and Held (2005) for more details). Integrated Nested Laplace Approximation will be considered in order to fit this kind of models. The analysis will be performed by using covariates such as age, gender, study level, Gini index, incomes, number of bars and regulations/sanctions.

Nonparametric level set estimation techniques will be applied in order to identify the hot-spots of emergency calls. Significant covariates detected from hierachical models fittings will be taken into account. In particular, differences between patterns by gender will be studied.

Keywords: Alcohol, bayesian hierarchical models, nonparametric level set estimation.

REFERENCES

Besag J., York J., Mollìe A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43(1)**, 1-20.

Rue H., Held L. (2005). Gaussian Markov Random Fields. London: Chapman and Hall, CRC Press.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

Casos prácticos da aplicación da estatística en enxeñaría naval e tecnoloxía mariña

Salvador Naya[1], Javier Tarrío-Saavedra[1], Luis Carral[2]

[1]CITIC, Grupo MODES, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña
[2]Campus Industrial de Ferrol, Grupo de Enxeñaría Mixto, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña

RESUMO

O uso da estatística é fundamental en todas as áreas da industria, máis se cabe no marco da actual cuarta revolución industrial, que implica a súa dixitalización a través da aplicación de tecnoloxías como a robótica, a simulación, a realidade aumentada ou a analítica de datos. Precisamente, dentro da analítica de datos, encádranse as presentes aplicacións da estatística para a resolución de casos reais de estudo na construción e deseño de buques. De feito, neste traballo presentaranse as técnicas e procedementos estatísticos utilizados para o control do tempo de tránsito a través do Canal de Panamá Ampliado, facendo posible a detección de patróns de aprendizaxe, ademais de para a medida da complexidade dos proxectos de buques en estaleiro e para o deseño de barcos. Doutra banda, desde a perspectiva da sustentabilidade, describirase o uso de técnicas estatísticas para a selección de materiais a utilizar para a construción de arrecifes artificiais no ámbito específico das rías galegas e cuxa finalidade última é a rexeneración da biodiversidade mariña. A aplicación do deseño de experimentos, gráficos de control para variables, modelos de regresión, análise de redes e métodos de clasificación non supervisada, entre outros, é fundamental para a consecución dos obxectivos mencionados. Se usará o paquete en `R` `qcr`.

Palabras e frases chave: Arrecifes Artificiais, Canle de Panamá, Construción Naval, Control Estatístico da Calidade, Aprendizaxe Estatística.

1. CONTROL DO TEMPO DE TRÁNSITO DE BUQUES NO CANAL DE PANAMÁ

O Canal de Panamá Ampliado (EPC) ven de inaugurarse o 26 de xuño de 2016, á fin de adaptarse ao maior tamaño dos buques Neo-panamax e, por tanto, permitir o seu tránsito a través do istmo de Panamá, aforrando tempo e outros recursos da navieiras, ademais de axudar a sustentar a economía do país. Como en toda instalación de nova creación, é fundamental monitorizar e controlar a calidade do servizo que se presta. Neste caso en particular, o obxectivo é a identificación de posibles patróns de aprendizaxe nas instalacións e pilotos do EPC, resultado das accións de formación, adestramento e melloras nas instalacións, ademais da práctica adquirida. Tendo en conta que a variable crítica para a calidade do proceso de tránsito de buques a través das esclusas é o tempo de tránsito, para tal fin, pódense aplicar gráficos de control tradicionais Shewhart, unha vez verificado o cumprimento das hipóteses de partida (observacións normais e independentes). En concreto, dada a dificultade de agrupar observacións en grupos homoxéneos, poden aplicarse os gráficos de control de medidas individuais. Previamente, hai que ter en conta que variables como as esclusas (Agua Clara, no Atlántico, e Cocolí, no Pacífico), a dirección de tránsito (norte ou sur) e o tipo de buque (sendo os máis comúns os portacontenedores, de trasporte de gas natural licuado ou LNG, e de transporte de gas licuado do petróleo ou LPG) inflúen significativamente no tempo de tránsito (Carral et al., 2020, 2021). A Figura mostra un exemplo de aplicación dos gráficos de control de medidas individuais aos datos de tempo de tránsito de buques portacontenedores con dirección sur a través da esclusa Agua Clara. Mediante os histogramas da dereita se observa que os tempos de tránsito correspondentes ao período 2017-2019 tenden a ser menores en

media e en dispersión se estes se comparan cos tempos de tránsito correspondentes ao primeiro mes de apertura do EPC. Este cambio pode estar relacionado coa existencia dun patrón de aprendizaxe, que se ve máis claramente nos gráficos de control da dereita da Figura 1. No panel superior dereito se amosa o gráfico de control de medidas individuais construído a partir dos tempos de tránsito dos buques que cruzan no primeiro mes. Unha vez estimados os límites de control con esta mostra de calibrado (Fase I), se procede a monitorizar o resto de tránsitos (Fase II) no panel inferior dereito. Como resultado, identifícase un patrón de cambio paulatino de nivel (observado a través da identificación de rachas de máis de seis observacións a un lado da liña central, en amarelo, e tamén valores fora do límite de control inferior, en vermello). O proceso comeza a estar fora de control, con respecto aos tránsitos do primeiro mes, a partir do 16º tránsito da mostra de monitorizado, tránsito a partir do cal se identifica o resultado do adestramento e formación levada a cabo pola Autoridade do Canal de Panamá, ademais de pola experiencia acumulada.

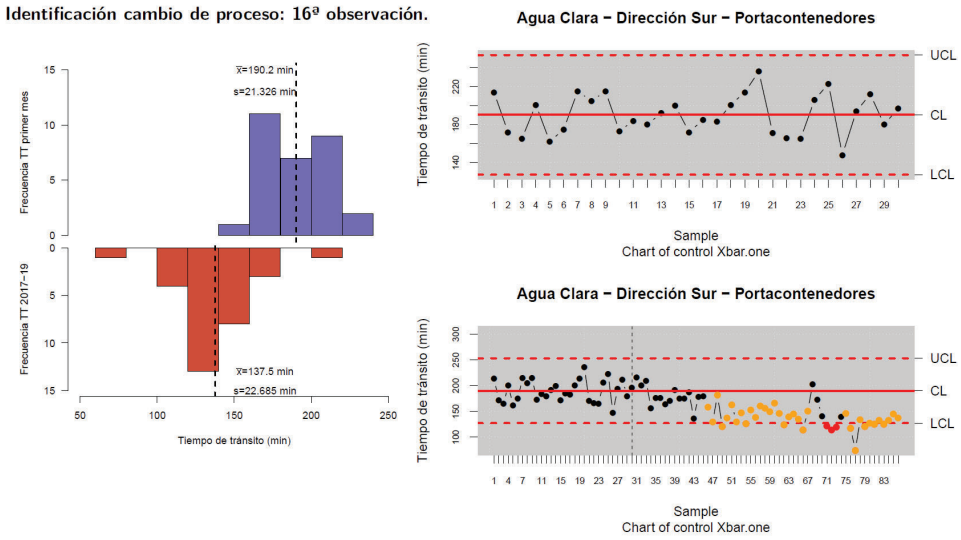


Figura 1: Histogramas e gráficos de control para medidas individuais correspondentes ao tempo de tránsito dos buques portacontenedores a través da esclusa de Agua Clara, no Canal Ampliado de Panamá.

Neste traballo, tamén se amosará a aplicación de gráficos de control multivariante T^2 de Hotelling, tendo en conta que cada tránsito dun buque a través do EPC está caracterizado polo tránsito a través de dúas esclusas, ademais de avaliar en que grao as instalacións, pilotos e empregados en xeral cumplen as especificacións que marca a Autoridade do Canal de Panamá, a partires da aplicación de técnicas de análise de capacidade de procesos (Montgomery, 2019).

2. ELECCIÓN DE MATERIAIS PARA A CONSTRUCCIÓN DE ARRECIFES ARTIFICIAIS

Por outra banda, neste traballo preséntase tamén o caso do desenvolvemento de arrecifes artificiais para a rexeneración da fauna mariña nas rías galegas (Carral et al., 2018, 2023a, 2023b). En concreto, se presentarán os resultados obtidos en recentes traballos, relacionados coa selección de materiais para a elaboración dos arrecifes.

Por un lado, faise un estudo para avaliar a posibilidade da substitución do árido do formigón por residuos da industria conserveira galega, como son as cunchas de diferentes bivalvos entre os que se atopan as ameixas, as ostras e as vieiras. A través de técnicas de análise da varianza e o axuste de modelos GAM e lineais multivariantes identifícase os efectos que sobre a resistencia á compresión do formigón teñen variables relacionadas coa cantidade de calcita (variable dependendo da especia de bivalvo), a procedencia desta (a partir de vieira ou ostra) e o tamaño dos áridos obtidos das cunchas. Como conclusión, amosarase que os parámetros de deseño do novo formigón para elaborar arrecifes artificiais deberían corresponderse a un tamaño de árido medio e unha porcentaxe en peso en calcita alta, sendo aconsellable que se obteña a partir de cunchas de ostra (Carral et al., 2023a).

Por outro lado, tamén se presenta un estudo sobre a durabilidade do formigón a utilizar como arrecife artificial, definindo como variables indicadoras a cantidade de auga absorbida, o volume perdido e a resistencia a compresión en ensaios de vida acelerada (Carral et al., 2023b).

REFERENCIAS

Carral, L., Alvarez-Feal, J. C., Tarrio-Saavedra, J., Guerreiro, M. J. R., Fraguela, J. Á. (2018). Social interest in developing a green modular artificial reef structure in concrete for the ecosystems of the Galician rías. Journal of Cleaner Production, 172, 1881-1898.

Carral, L., Tarrío-Saavedra, J., Álvarez-Feal, J. C., Naya, S., Sabonge, R. (2020). Modeling and forecasting of Neopanamax vessel transit time for traffic management in the Panama Canal. Journal of Marine Science and Technology, 25, 379-396.

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

Carral, L., Lamas-Galdo, M. I., Buenhombre, J. L. M., Barros, J. J. C., Naya, S., Tarrio-Saavedra, J. (2023a). Application of residuals from purification of bivalve molluscs in Galician to facilitate marine ecosystem resiliency through artificial reefs with shells–One generation. Science of the Total Environment, 856, 159095.

Carral, L., Tarrío-Saavedra, J., Barros, J. J. C., Fabal, C. C., Ramil, A., Álvarez-Feal, C. (2023b). Considerations on the programmed functional life (one generation) of a green artificial reef in terms of the sustainability of the modified ecosystem. Heliyon, 9(4).

Montgomery, D. C. (2019). Introduction to statistical quality control. John wiley & sons.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Stochastic orders and ageing properties tests

Idir Arab[1], Tommaso Lando[2] and Paulo Eduardo Oliveira[1]

[1]CMUC, Dep. Mathematics, Univ. Coimbra, Portugal
[2]Dep. Economics, Univ. Bergamo, Italy, and Dep. Finance, VŠB-TU Ostrava, Czech Republic

ABSTRACT

The study of hazard rate properties is a popular way for benchmarking ageing properties. Indeed, as the exponential distribution is memoryless, or does not show ageing effects, and has a constant hazard rate, the monotonicity of the hazard rate may be taken as a manifestation of negative or positive ageing effects. Hence the interest on testing about the increasingness of the hazard rate or the increasingness in average, corresponding to the well known IHR or IHRA classes of distributions, respectively. Tests have been proposed, mostly with justifications based on asymptotic properties. We are interested in finite sample, and especially, small sample properties. We consider test statistics that show monotonicity properties with respect to suitable stochastic ordering notions, exploring these to prove good testing behaviour, regardless of the sample size. This class test statistics is shown to be large enough, including, for example, the popular Kolmogorov-Smirnov statistic.

However, the IHR class is relatively restrictive, as it implies the existence of moments of every order. So, heavy-tailed distributions, for example, are excluded from the scope of applicability of the previous approach. Noting that the log-logistic distribution satisfies a multiplicative form of the lack of memory property, we propose to benchmark against the log-logistic distribution. The ordering thus obtained is now based on monotonicity properties of the odds ratio. We show that a large class of estimators preserves this monotonicity, hence allowing to derive small sample properties for goodness-of-fit test based on them.

Some numerical examples giving some insight about the performance of the estimators and tests considered are included.

Keywords: hazard rate, odds rate, heavy tail, monotone transformations, nonparametric tests.

REFERENCES

Groeneboom, P., Jongbloed, G. (2014) Nonparametric estimation under shape constraints. Cambridge University Press.

Hall, P., Van Keilegom, I. (2005) Testing for monotone increasing hazard rate. Annals of Statistics, 33, 1109–1137. doi:10.1214/009053605000000039.

Lando, T., Arab, I., and Oliveira, P.E. (2022) Properties of increasing odds rate distributions with a statistical application. Journal of Statistical Planning and Inference, 221, 313–325, doi:10.1016/j.jspi.2022.05.004.

Lando, T., Arab, I., and Oliveira, P.E. (2023a) Transform orders and stochastic monotonicity of statistical functionals. Scandinavian Journal of Statistics, 50, 1183–1200, doi:10.1111/sjos.12629.

Lando, T., Arab, I., and Oliveira, P.E. (2023b) Nonparametric inference about increasing odds rate distributions. Journal of Nonparametric Statistics (to appear), doi:10.1080/10485252.2023.2220050.

Shaked, M., Shanthikumar, J.G. (2007) Stochastic orders. Springer.

Tenga, R., Santner, T.J. (1984) Testing goodness of fit to the increasing failure rate family. Naval Research Logistics Quarterly, 31, 617–630, doi:10.1002/nav.3800310411.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Smooth $k-$sample tests under left truncation

Adrián Lago[1] Ingrid Van Keilegom[2] Jacobo de Uña Álvarez[1] and Juan Carlos Pardo Fernández[3]

[1]CINBIO, Universidade de Vigo, Department of Statistics and Operations Research, SiDOR research group, Vigo, Spain.

[2]KU Leuven, Department of Decision Sciences and Information Management, ORSTAT reseach group, Leuven, Belgium.

[3]CITMAGA, Universidade de Vigo, Department of Statistics and Operations Research, SiDOR research group, Vigo, Spain.

ABSTRACT

Truncation is a phenomenon that causes bias when estimating population quantities, such as the survival function. To overcome such a problem, the Lynden-Bell (1971) estimator could be employed. Following from it, one can also define an estimator of the density function, as seen in Zhou (1999). On the other hand, the comparison of the distribution of target variables in k independent populations is of great importance in different studies or experiments. As such, the development of proper techniques to do so with truncated data seems necessary. In this work, we propose a test depending on the density estimator for truncated data. The test statistic is based on a quadratic distance between the estimator in every subsample and the one of the pooled sample. Its asymptotic null distribution is studied and, due to the difficulty of application in practice, a bootstrap resampling plan is proposed to approximate the null distribution of the test statistic. The performance of this method is studied via Monte Carlo simulations. The choice of the bandwidth plays a central role on the performance of the test, as its value affects the power, thus it is carefully studied. The proposed test is compared to other existing tests for left-truncated data, such as the log-rank and the Kolmogorov-Smirnov. Lastly, data regarding abortion times will be employed to exemplify the performance of the test.

Keywords: Survival Analysis, $k-$sample problem, truncation, bootstrap

REFERENCES

Lynden-Bell, D. (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. Monthly Notices of the Royal Astronomical Society, 155, 95-118.
Zhou, Y. (1999) Asymptotic representations for kernel density and hazard function estimators with left truncation. Statistica Sinica, 9, 521-533.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Contrastes de comparación de procesos puntuais sobre grafos lineares

M. I. Borrajo[1], I. González-Pérez[2] e W. González-Manteiga[1]

[1] CITMAga. Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.
[2] ETH Zürich.

RESUMO

Os conxuntos de datos que representan a localización espacial dunha serie de observacións aparecen nunha gran variedade de escenarios, por exemplo, árbores nun bosque, incendios forestais nunha certa rexión ou accidentes de tráfico nunha rede de estradas. Este último é un exemplo de patróns puntuais que non se atopan nunha subrexión bidimensional do plano euclidiano, senón que están restrinxidos a un subconxunto unidimensional do mesmo. Diremos que este tipo de patróns se atopan sobre un grafo linear. Analizar procesos puntuais en grafos lineares presenta maiores complexidades que traballar en calquera espazo euclidiano, debido principalmente á métrica asociada.

Un problema moi estudado en estatística é a comparación de poboacións, é dicir, determinar se dúas (ou máis) mostras son xeradas polo mesmo proceso estocástico. Este problema tamén xorde cando se trata de procesos puntuais, por exemplo, a distribución de dúas especies de flora nun bosque, puntos de ignición de incendios forestais naturais ou provocados, colisións coche-coche e coche-moto nunha rede viaria...

No eido dos procesos puntuais espaciais, xa se ten abordado este problema de comparación, porén este non é o caso dos procesos puntuais noutros dominios, como por exemplo os grafos lineares. Neste traballo estudamos o problema de comparación de dúas mostras para procesos puntuais en grafos lineares, propoñendo varios métodos de contrastes específicos: un estatístico de tipo Kolmogorov-Smirnov, outro de tipo Cramér-von-Mises e un terceiro baseado no risco relativo. Realízase un exhaustivo estudo de simulación para detallar o rendemento das nosas propostas sobre mostras finitas, e tamén se aplican estas novas ideas a un conxunto de datos sobre colisións de tráfico en Río de Janeiro (Brasil).

Palabras e frases chave: procesos puntuais; grafos lineares; comparación de poboacións.

REFERENCIAS

McSwiggan, G., Baddeley, A. e Nair, G. (2017) Kernel density estimation on a linear network. Scandinavian Journal of Statistics, 44, 324–345.

Zhang, T. e Zhuang, R. (2017) Testing proportionality between the first-order intensity functions of spatial point processes. Journal of Multivariate Analysis, 155, 78–82.

Okabe, A., Satoh, T e Sugihara, K. (2009) A kernel density estimation method for networks, its computational method and a GIS-based tool. International Journal of Geographical Information Science, 23, 7–32.

Okabe, A. e Sugihara, K. (2012) Spatial analysis along networks: statistical and computational methods. John Wiley and Sons.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Avances en contrastes de bondade de axuste baseados en estatísticos de enerxía para datos censurados

María Vidal-García[1], Rosa M. Crujeiras[1] e Wenceslao González-Manteiga[1]

[1]CITMAga, 15782 Santiago de Compostela, Spain. Department of Statistics, Mathematical Analysis and Optimization. Universidade de Santiago de Compostela (USC).

RESUMO

Dende a súa introdución por Székely et al. (2007), a correlación de distancias e máis en xeral os estatísticos de enerxía amosan numerosas aplicacións á hora de construir contrastes de especificación, por exemplo en tests de independencia (Székely et al., 2007), no problema de k mostras (Székely e Rizzo, 2004) ou en contrastes de bondade de axuste (Székely e Rizzo, 2017). Estes estatísticos están baseados fundamentalmente no cálculo da distancia de enerxías (*energy distance*) entre distribucións. A principal vantaxe desta magnitude é que, cando se calcula a distancia de enerxías entre funcións de distribución empíricas, a expresión resultante pódese escribir en termos de momentos de distancias entre observacións, o que facilita o seu cálculo. Esta expresión conecta ademais coa teoría dos U-estatísticos facilitando o estudo das súas propiedades.

Na área de Análise de Supervivencia, a adaptación deste tipo de ferramentas comezou a explorarse máis tarde con traballos como Fernández e Gretton (2019) ou Edelmann et al. (2022).

Nesta charla introduciremos os estatísticos de enerxía, revisaremos as súas aplicacións en contrastes de bondade de axuste e veremos como adaptar estes tests á presenza de censura aleatoria pola dereita en contextos tanto de hipótese nula simple como composta. Revisaremos as vantaxes e limitacións de cada proposta e exploraremos posibles solucións. Por último, ilustraremos o comportamento das ferramentas presentadas mediante simulacións.

Palabras e frases chave: energy distance, contraste de bondade de axuste, censura, análise de supervivencia

REFERENCIAS

Edelmann, D., Welchowski, T., e Benner, A. (2022) A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. Biometrics, 78(3), 867-879.

Fernández, T. e Gretton, A. (2019) A maximum-mean-discrepancy goodness-of-fit test for censored data. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 2966-2975). PMLR.

Székely, G. J. e Rizzo, M. L. (2004) Testing for equal distributions in high dimension. InterStat, 5(16.10), 1249-1272.

Székely, G. J. e Rizzo, M. L. (2017) The energy of data. Annual Review of Statistics and Its Application, 4, 447-479.

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

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Predicting intensive care unit bed occupancy under random regression coefficient Poisson models: Application to the COVID-19 pandemic in Galicia

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

[1]Universidade da Coruña, CITIC, Spain.
[2]Universidade da Coruña, CITIC, Spain.
[3]Universidade Miguel Hernández de Elche, IUCIO, Spain.

ABSTRACT

COVID-19 has highlighted the need for spatio-temporal monitoring of care capacity. In this work, we develop methodology in Small Area Estimation to derive predictors of occupied Intensive Care Unit beds under a random regression coefficient mixed model, introducing bootstrap estimators of mean squared errors. Maximum likelihood estimators of model parameters and random effects mode predictors are calculated using a Laplace approximation algorithm. Simulation experiments are conducted to investigate the behaviour of the fitting algorithm, the predictor and the mean squared error estimator. The new statistical methodology is applied to map COVID-19 daily occupancy in intensive care units of the seven Galician health areas between November 2020 and March 2022. The results reveal the potential of the predictors to support health planning, both in low and high care pressure scenarios, as they show optimal accuracy for public use in a wide variety of conditions, such as different variants of SARS-CoV-2, restrictive measures or different levels of vaccination. In fact, the defined methodology is being programmed in an interactive Shiny application to facilitate the management of healthcare resources at user level.

Keywords: Small Area Estimation, Random regression coefficient mixed models, Bootstrap, Intensive Care Unit occupancy, COVID-19.

1. INTRODUCTION

COVID-19, a disease caused by the SARS-CoV-2 virus, has had disastrous consequences in all key areas of human well-being. Since the first cases were reported in December 2019 in Hubei Province in the People's Republic of China, the contagion has spread to all corners of the world at a rate of spread far beyond the capacity of health systems to provide care.

In fact, the saturation of hospitals and especially of Intensive Care Units (ICUs) has been one of the major bottlenecks to survival. This situation, coupled with the predicted emergence of new epidemics in the not too distant future, highlights the need to develop predictive tools that allow precise temporal and spatial monitoring of the saturation of care capacity.

In this line, since 2020, there has been a succession of investigations, mainly involving inference on the incidence rate and on the length of stay (LoS) in hospital. Many of the approaches are based on models traditionally used in epidemiology such as compartmental models, like SIR and SEIR, which categorise and assign individuals into pre-defined compartments, modelling their transition from one to the other (Susceptible - Infected - Recovered for SIR and Susceptible - Exposed - Infected - Recovered for SEIR). However, their use for COVID-19 pandemic modelling presents multiple difficulties. These models are based on a set of differential equations with various initial conditions and adaptive parameters, the values of which are usually only accurately obtained when pandemics are overcome (García-Vicuña et al., 2022; Ibáñez et al., 2021). Thus, the availability of daily data on new positives, as well as their quality throughout the pandemic, has been problematic. In addition to the lack of knowledge about the virus in the first waves and

the overburdening of health resources that resulted in totally insufficient diagnostic tests and the consequent underestimation of those infected, there have also been changes in the criteria used to collect and disseminate the figures as the pandemic and knowledge about the virus have evolved.

On the other hand, experts in the field of survival analysis have also been actively involved. Notably, López-Cheda et al. (2021) propose the prediction of the number of beds needed for COVID-19 patients through the estimation of the lengths of stay (LoS) of these patients developing a non-parametric mixture-type cure model adjusted for key factors such as age and sex. The results show optimal performance, both in simulation and in application to real data from the Autonomous Community of Galicia in the period between March and May 2020, outperforming other classical estimators in the field when modelling events that are not suffered by all individuals. Safari et al. (2022), develop an estimator of the probability of cure in a mixture cure model when some individuals are known to be cured and apply it to the estimation of the event of interest, ICU admission, in data from the Autonomous Community of Galicia between March and May 2020, illustrating the good theoretical and applicability properties of the model. Given these optimal results, García-Vicuña et al. (2023) incorporate these estimators into discrete-event simulation models to define the key floor and ICU LoS parameters to properly define the trajectory of patients through the hospital from admission to discharge or death and, therefore, to make a prediction of the beds needed during the course of successive waves of the COVID-19 pandemic.

Despite these magnificent contributions, there are still major challenges to be overcome in the estimation and prediction of hospital saturation. As the authors point out, the dependence on the existence of data and its quality is key, being a handicap during this pandemic, where delays in reporting, changes in registration criteria and a marked spatial-temporal heterogeneity of the disease stand out. In this respect, it should be noted that COVID-19 has not affected the different geographical areas of each country in the same way, neither in intensity nor in speed, due to a variety of factors that have a different influence on its spread across the different territories. In fact, the Instituto de Salud Carlos III (ISCIII) has funded the COV20-00881 project to identify these factors and be able to act on them to prevent future outbreaks and their intensity (Centro Nacional de Epidemiología (CNE), 2022).

In order to incorporate this heterogeneity in modelling and to obtain indicators for planning hospital resources and intervention measures, analyses are required at a lower level of aggregation than country, community or province, with health areas being potentially interesting spatial target units as shown by their prominence during the return to normality phase according to the results of the risk indicators in their demarcation.

This poses a challenge for obtaining accurate estimates, which can be addressed by Small Area Estimation (SAE), a field of statistics that uses techniques such as mixed models to address these problems. The SAE methodology responds to the need to produce precise estimates of indicators of interest in areas or domains, geographic or otherwise, e.g. demographic, where the number of observations is too small or even zero to obtain precise estimates. Such a line of research started with the study by Fay and Herriot (1979), and since then, publications of inference on continuous and discrete variables at the domain level have followed one after the other. For a detailed review of this methodology, see articles and monographs such as Pfeffermann (2013), Rao and Molina (2015), Morales et al. (2021), among others.

However, to date, few studies have applied the SAE methodology to the study of the COVID-19 pandemic, such as the study by Martínez-Beneito et al. (2022), which aims to draw inferences about the incidence of COVID-19 cases. In this research, following the line of mixed models with spatio-temporal dependence of random effects, they try to achieve greater flexibility by also incorporating random slopes, being, to date, to the best of our knowledge, the only study on COVID-19 and SAE with this type of model. Simulation analysis and application with real data from the Community of Valencia and Castilla León demonstrate their potential usefulness in detecting different patterns of evolution depending on the geographical area, facilitating the most appropriate decision making for each territory in terms of epidemic containment.

In this work, for the first time, we derive predictors of occupied ICU beds counts and occupancy ratios under an area random regression coefficient Poisson model (ARRCP model), that was first defined in Diz-Rosales et al. (2023), introducing bootstrap estimators of mean squared error. Maximum likelihood estimator of model parameters and mode predictors of random effects are calculated using a Laplace approximation algorithm. In addition, we empirically investigate

the behaviour of the fitting algorithm and the MSE estimator through simulation studies, before proceeding to its application to real data, being the main objective to predict the counts and proportions of ICU beds occupied by COVID-19 patients by health area and day in Galicia, between November 2020 and March 2022, to support health planning.

The summary is structured as follows. Section 2 describes the dataset. Section 3 presents the ARRCP model and gives the basic mathematical developments of the predictor chosen for application to real data, the plug-in predictor. Section 4 applies the model to data from the seven health areas of Galicia between November 2020 and May 2021 to predict the number of daily ICU beds occupied by COVID-19 patients. Given the good results obtained during the modelling, Section 5 presents the application of the proposed methodology to make future predictions of COVID occupancy in ICU, between May 2021 and March 2022, in time horizons of 3, 5 and 7 days. The summary ends with the main conclusions and a series of references.

2. DATA DESCRIPTION

It is worth starting the description of the dataset by highlighting that we have built the database from scratch by collecting information between November 2020 and March 2022 for each of the seven existing health areas in Galicia: 1) Health Area of A Coruña e Cee; 2) Health Area of Ferrol; 3) Health Area of Lugo, A Mariña e Monforte de Lemos; 4) Health Area of Ourense, Verín e O Barco de Valdeorras, 5) Health Area of Pontevedra e o Salnés; 6) Health Area of Santiago de Compostela e a Barbanza; and 7) Health Area of Vigo.

In this way, we collect information corresponding to 181 consecutive days for each health area, between 2nd November 2020 and 2nd May 2021, which is used to fit the ARRCP model, and information corresponding to 307 consecutive days for each health area, between 3rd May 2021 and 6th March 2022, with which we evaluate the predictive quality with data not used in the fit.

At this point, it is also important to highlight the reason for the selection of the time range. The starting date of the model adjustment is marked by the availability of the data to be used relating to the occupancy of ICU beds and ward beds by COVID-19 patients, as we have records starting on 7th October 2020. Therefore, we decided to leave a time margin and start on 2nd November 2020 to incorporate indicator variables and their time-delayed versions into the study process. In this way, we aim to incorporate the dynamic of delayed data reporting that has been experienced throughout the pandemic and the dynamics of the disease itself, with lags between infection, admission to the ward, ICU or unfortunately death. The end date of 6 March 2022 allows for a highly variable scenario to be taken into account, as six epidemic periods (Ministerio de Sanidad, 2022) occur in these 488 days, with scenarios of both high and low healthcare pressure, incorporating the effect of the transition of three variants of the SARS-CoV-2: Alpha, Delta and Omicron, as illustrated by viral load analyses in wastewater from Galicia (Trigo-Tasende et al., 2023).

As a result, the dataset constructed has a total of 3416 rows, corresponding to the observation of 488 days for each of the seven health areas. For each area and day, the target variable of the ARRCP model is the count of people hospitalised with COVID-19 in ICU and we assessed the inclusion of the following auxiliary variables that were used by health authorities in monitoring the level of risk or analysed in the reviewed literature (Ministerio de Sanidad, 2022):

- Variables for the assessment of the level of transmission: CI14 (14-day cumulative incidence defined as the number of confirmed 14-day COVID-19 cases per 100,000 inhabitants); CI7 (7-day cumulative incidence defined as the number of confirmed COVID-19 cases in 7 days per 100,000 inhabitants); and positive.rate7 (percentage of the number of weekly tests with a positive result compared to the number of tests performed in that period).
- Variables for the assessment of the level of utilisation of care services by COVID-19: ward.rate (number of hospital beds occupied by COVID-19 cases per 100,000 inhabitants); disch.rate14 (number of COVID-19 discharges within 14 days per 100,000 inhabitants); and disch.rate7 (number of COVID-19 discharges within 7 days per 100,000 inhabitants).
- Variables for assessing the level of severity: acute.rate7 (percentage of cases admitted to ICU among all COVID-19 inpatients within seven days); death.rate7 (7-day rate of COVID-19 deaths per 1,000,000 inhabitants); and death.rate14 (14-day rate of COVID-19 deaths per 1,000,000 inhabitants).

For the construction of almost all auxiliary variables, we used open data from the Servizo Galego de Saúde (SERGAS, 2023), while to obtain the target variable, the ward.rate and acute.rate7 variables, we resorted, after checking their veracity with information from official sources, to an open public repository in Github that stores the official publications of hospital occupancy that SERGAS published daily during the pandemic (Lipido, 2023).

Finally, it should also be noted that due to the interest in knowing the population proportion of ICU beds occupied by COVID-19 patients, and in line with other works in SAE, such as Martínez-Beneito et al. (2022), the model, defined in Section 3, has an offset, which in this case corresponds to the population living in each health area in 2021 according to the figures of the Population and Housing Census of 2021, published by the Instituto Nacional de Estadística (INE, 2021).

3. THE ARRCP MODEL

Let us consider a count variable y_{it} taking values on $\mathbb{N} \cup \{0\}$, where $i \in \mathbb{I} = \{1, \ldots, I\}$ and $t \in \mathbb{T} = \{1, \ldots, T\}$. Let $D = IT$ be the total number of y-values. For example, y_{it} could be the number of people with COVID-19 disease hospitalized in an Intensive Care Unit (ICU), the indexes i and t may represent health area and day, and D is the total number of domains defined by the crossings of the variables health area and day. In other words, we have a region partitioned into health areas and days. We further assume that each day can be grouped into one, and only one, of the W clusters, $\mathbb{T}_1, \ldots, \mathbb{T}_W$, of time window variable. Let $w(t)$ be the number of the category to which day t belongs, so that $w(t) \in \mathbb{W} = \{1, \ldots, W\}$. The number of days in the cluster $\mathbb{T}_w$ is $m_w = \#(\mathbb{T}_w)$, so that $D = I \sum_{w=1}^{W} m_w$.

We are dealing with area-level data for modelling and predicting the target variable y_{it}. Let us assume that we have p explanatory variables with values $x_{\ell,it}$, $\ell \in \mathbb{P} = \{1, \ldots, p\}$, $i \in \mathbb{I}$, $t \in \mathbb{T}$. For models with intercept, we take $x_{0,it} = 1$ for all i and t. In what follows, we present a area level random regression coefficient Poisson (ARRCP) model.

Let u_{iw}, $i \in \mathbb{I}$, $w \in \mathbb{W}$ be i.i.d. $N(0,1)$ random variables. Let $\phi_\ell > 0$, $\ell \in \mathbb{P}$, be unknown standard deviation parameters. Let $\rho_{rs} \in (-1,1)$, $r < s$, $r, s \in \mathbb{P}$, be unknown correlation parameters. Let $\boldsymbol{v}_i = (v_{1,i}, \ldots, v_{p,i})'$, $i \in \mathbb{I}$, be i.i.d. random vectors such that

$$\underset{1\le\ell\le p}{\text{diag}}\,(\phi_\ell)\boldsymbol{v}_i \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{vi}^{\phi\rho}), \quad \boldsymbol{V}_{vi}^{\phi\rho} = \begin{pmatrix} \phi_1^2 & \phi_1\phi_2\rho_{12} & \ldots & \phi_1\phi_p\rho_{1p} \\ \phi_2\phi_1\rho_{12} & \phi_2^2 & \ldots & \phi_2\phi_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_p\phi_1\rho_{1p} & \phi_p\phi_2\rho_{2p} & \ldots & \phi_p^2 \end{pmatrix}.$$

Therefore, the variance of $\boldsymbol{v}_i$ is

$$\boldsymbol{V}_{vi} = \text{var}(\boldsymbol{v}_i) = \underset{1\le\ell\le p}{\text{diag}}\,(\phi_\ell^{-1})\boldsymbol{V}_{vi}^{\phi\rho}\underset{1\le\ell\le p}{\text{diag}}\,(\phi_\ell^{-1}) = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{12} & 1 & \ldots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \ldots & 1 \end{pmatrix}.$$

Let us define the vectors

$$\boldsymbol{u} = \underset{1\le i\le I}{\text{col}}\,(\underset{1\le w\le W}{\text{col}}\,(u_{iw})) \sim N_{IW}(\boldsymbol{0}, \boldsymbol{I}_{IW}), \quad \boldsymbol{v} = \underset{1\le i\le I}{\text{col}}\,(\boldsymbol{v}_i) \sim N_{pI}(\boldsymbol{0}, \boldsymbol{V}_v),$$

where $\boldsymbol{V}_v = \underset{1\le i\le I}{\text{diag}}\,(\boldsymbol{V}_{vi})$ and where diag and col are the diagonal and the column operator, respectively. We assume that $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent. The distribution of the target variable y_{it}, conditioned to the random effects $u_{iw(t)}$, $v_{\ell,i}$, $\ell \in \mathbb{P}$, is

$$y_{it}|_{u_{iw(t)}, v_{1,i}, \ldots, v_{p,i}} \sim \text{Poisson}(\nu_{it}p_{it}), \quad i \in \mathbb{I},\ t \in \mathbb{T},$$

where the offset (or size) parameters $\nu_{it} > 0$ are known and correspond to the population size when the model is applied to real data, and the binomial probability, p_{it}, is the target parameter with range (0,1). For the natural parameters, we assume

$$\eta_{it} = \log\mu_{it} = \log\nu_{it} + \log p_{it} = \log\nu_{it} + \sum_{\ell=1}^{p}\beta_\ell x_{\ell,it} + \sigma u_{iw(t)} + \sum_{\ell=1}^{p}\phi_\ell v_{\ell,i}x_{\ell,it}, \ \ i \in \mathbb{I},\, t \in \mathbb{T}, \quad (1)$$

where $\mu_{it} = E[y_{it}|u_{iw(t)}, v_{1,i}, \ldots, v_{p,i}]$. We may write $\boldsymbol{x}_{it}\boldsymbol{\beta} = \sum_{\ell=1}^{p} \beta_\ell x_{\ell,it}$, where $\boldsymbol{\beta} = \underset{1\leq\ell\leq p}{\text{col}}\,(\beta_\ell)$ is the column vector of regression parameters and $\boldsymbol{x}_{it} = \underset{1\leq\ell\leq p}{\text{col}'}\,(x_{\ell,it})$ is the row vector of known auxiliary variables. To finish the definition of the ARRCP model, we assume that the y_{it}'s are independent conditioned to $\boldsymbol{u}$ and $\boldsymbol{v}$. The variance component parameters are $\sigma > 0$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \in \mathbb{R}_+^p$ and $\boldsymbol{\rho} = (\rho_{12}, \ldots, \rho_{1p}, \ldots, \rho_{p-1p})' \in (-1, 1)^{p(p-1)/2}$, where $\mathbb{R}_+ = (0, \infty)$. The vector of model parameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma, \boldsymbol{\phi}', \boldsymbol{\rho}')'$. The total number of random effects is $H = IW + pI$.

With the ARRCP model defined, we proceed to carry out the maximization, deriving the maximum likelihood estimators of the model parameters, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, $\hat{\phi}$, and the mode predictors of the random effects, by means of a Laplace approximation algorithm (ML Laplace algorithm), using the R package lme4 1.1-33.

In addition, we define the best predictor (BP), the best empirical predictor (EBP) and the plug-in predictor (IN), to predict the ICU occupancy and ICU occupancy rate by COVID-19 patients. The best predictor (BP) of p_{it} is

$$\hat{p}_{it}^{bp} \quad = \quad \hat{p}_{it}^{bp}(\boldsymbol{\theta}) = E_\theta[p_{it}|\boldsymbol{y}] = E_\theta[p_{it}|\boldsymbol{y}_i]$$

The empirical best predictor (EBP) of p_{it} is $\hat{p}_{it}^{ebp} = \hat{p}_{it}^{bp}(\hat{\boldsymbol{\theta}})$. This predictor requires approximating a multivariate integral which we approximate by a Monte Carlo method.

The behaviour of these predictors is then evaluated by bootstrap simulation studies of at least 1000 iterations, comparing them with other predictors, such as the plug-in predictor (IN) of p_{it} and μ_{it} under the ARRCP model which has the form

$$\hat{p}_{it}^{in} = \exp\{\boldsymbol{x}_{it}\hat{\boldsymbol{\beta}} + \hat{\sigma}\hat{u}_{iw(t)} + \sum_{\ell=1}^{p} \hat{\phi}_\ell \hat{v}_{\ell,i} x_{\ell,it}\},$$

where $\hat{u}_{iw(t)}$ and $\hat{v}_{\ell,i}$, $\ell = 1, \ldots, p$, are the mode predictors taken from the output of the ML Laplace algorithm, being $\hat{\mu}_{it}^{in} = \nu_{it}\hat{p}_{it}^{in}$ the IN predictor of $\mu_{it} = \nu_{it}p_{it}$.

As a result of the simulations, the IN predictor shows the best computational efficiency-accuracy trade-off. To understand these results, it should be noted that the estimation of the BP and EBP is a very complex multivariate integral approximation problem. In order to perform this Monte Carlo approximation, we have generated simulations with different configurations in the number of replications, until we have tested, for the moment, the approximation with the generation of 2500 independent random variables. However, as we can see from the results, this number is insufficient. The BP and EBP incorporate the Monte Carlo variance, which is high, and which has a variance underlying the estimation of the multivariate integral, which would require substantially more than 2500 for an optimal approximation. By contrast, these simulations are highly computationally expensive, increasing simulation times to the order of days. Therefore, at the expense of obtaining more efficient results in terms of computational performance, and having verified the high performance of the IN predictor, we choose it as the starting predictor for this study.

Consequently, a second simulation study is performed to check the behaviour of the MSE estimator of this predictor based on the parametric bootstrap. As a result, a stabilisation of the performance metrics is obtained after 600 bootstrap iterations. The behaviour is practically unbiased, with values of relative root-MSE or $RRMSE_{it}$, lower than 25 %, calculated as the percentage of the root-MSE coefficient among a MSE taken as true in a simulation context. This, together with a feasible computational time, are optimal results in public statistics.

4. APPLICATION TO REAL DATA

During the model selection phase, we consider several criteria, highlighting that must be a model with easily accessible auxiliary information in epidemic contexts for its applicability to be viable. In addition, we considered: a) significance of model parameters and epidemiological interpretability; b) convergence of the ML-Laplace approximation algorithm; c) validity of model assumptions; and d) lower conditional AIC (for more informaiton on model selection in SAE see Vaida and Blanchard (2005) and Lombardía et al. (2017) among others).

As a result of the selection process, we define the ARRCP model, introduced in 1 in Section 3, with the following variables: y_{it} is the count of people with COVID-19 in ICUs by health area i

and day t, $\nu_{it} = N_{it}$ is the population size, $x_{0,it}$ is the intercept and $x_{1,it}$, $x_{2,it}$, $x_{3,it}$ are the values of the auxiliary variables acute.rate7, disch.rate14L3 and ward.rateL2 respectively reconverted to a population base of 1000 inhabitants to ensure the absence of convergence problems due to differences in scale. For clarification purposes, it should be noted that the suffixes L2 and L3 in the auxiliary variables refer to a time lag of 3 and 2 days with respect to the target variable, which falls within the confidence band indicated in the literature on disease dynamics (Ministerio de Sanidad, 2022). The selected model contains only one random slope for $x_{1,it}$. Thus, the natural parameter of the selected ARRCP model is as follows

$$\log \mu_{it} = \log \nu_{it} + \log p_{it} = \log \nu_{it} + \sum_{\ell=1}^{4} \beta_\ell x_{\ell,it} + \sigma u_{iw(t)} + \phi_1 v_{1,i} x_{1,it},$$

where $\nu_{it} = N_{it}$ is the population size in health area i and day t, $u_{iw(t)} \sim N(0,1)$, $(v_{1,i})' \sim N(\mathbf{0}, \boldsymbol{V}_1)$, $\mathbf{0} = (0,0)'$.

In the Table 1 we can contemplate the parameter estimates and the 95% CIs basic percentile bootstrap intervals, observing that they are all significant and that they exert a protective effect against the target variable, that is, when an increase in these variables produces an increase in the rate of ICU hospitalisation. This is coherent, since the fact that there are more patients will make it more likely that a portion of them will be admitted to the ICU. Along the same lines, when the weekly severity indicator increases, it is an indication that the infections evolve into severe cases. Even the fact that there are more discharges, which could be beneficial, should be seen in the light of the fact that more admissions have occurred previously.

	β_0	β_1	β_2	β_3	σ	ϕ_1
2.5%	-13.1571	4.1991	0.0659	2.9334	0.2847	0.2714
Estimate	-12.2635	5.0927	0.0934	3.2756	0.3237	0.6235
97.5%	-11.4101	5.9460	0.1212	3.6344	0.3679	1.2470

Table 1: Basic percentile confidence intervals α=5%.

Once the model is selected, it undergoes the diagnostic phase starting with the evaluation of the Pearson residuals, as visualised in Figure 1.

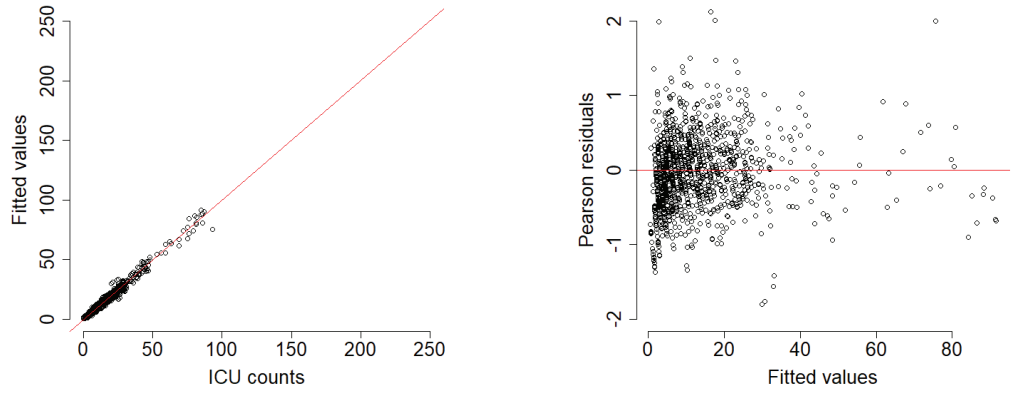


Figure 1: Diagnosis of the ARRCP model.

Following the good results, we proceed to further evaluate the performance in estimating the count of patients hospitalised in ICU due to COVID-19. For this purpose, in Figure 2 we plot the estimates of the ICU occupancy rate due to COVID-19 per 100,000 inhabitants per day and area.

As can be seen in Figure 2, the health areas that reached the highest ratio were the H.A. of A Coruña e Cee and the H.A. of Ferrol, at the peak in February 2021. It is interesting to note that in some health areas there are two large peaks, such as in the H.A. of Vigo and in the H.A. of

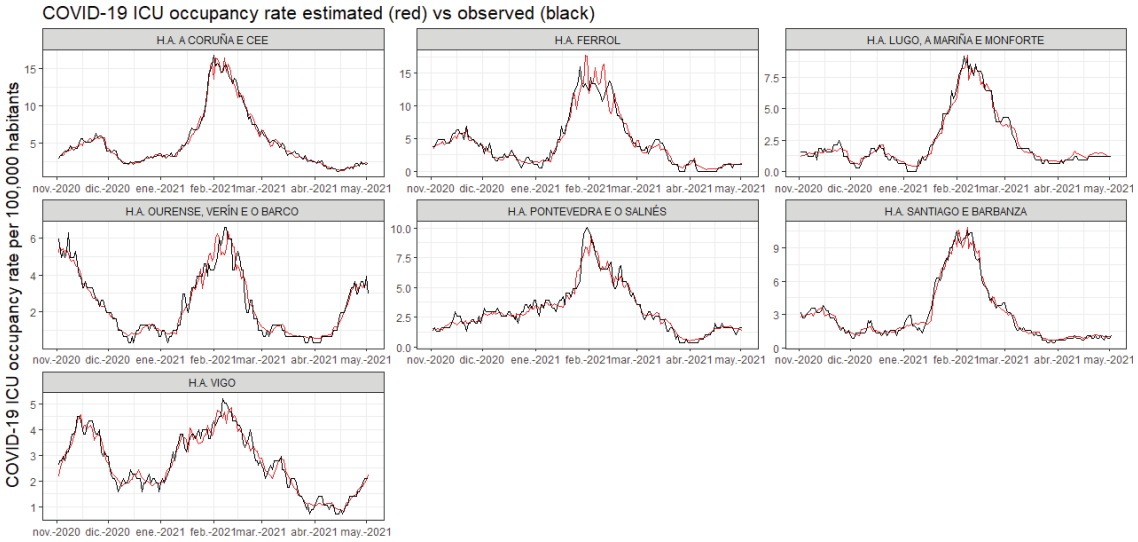


Figure 2: Observed (black line) vs estimated (red line) ICU occupancy rate per 100,000 inhabitants.

Ourense, Verín e o Barco de Valdeorras, and in the rest one. This illustrates the heterogeneity of the COVID-19 effect, where the acute peaks of care pressure are experienced at different stages.

In relation to the $RMSE$, the behaviour seems optimal in Figure 3, not exceeding on practically no occasion a deviation of 1 occupied ICU beds per 100,000 inhabitants per day, except in the H.A. of Ferrol, where higher values are reached between February and March 2021. In this period of time, in the rest of the health areas an increase in the RMSE is also detected, coinciding with a period of high healthcare pressure, so that in comparison, the increase seems to respond simply to the increase in the number of hospitalised patients.

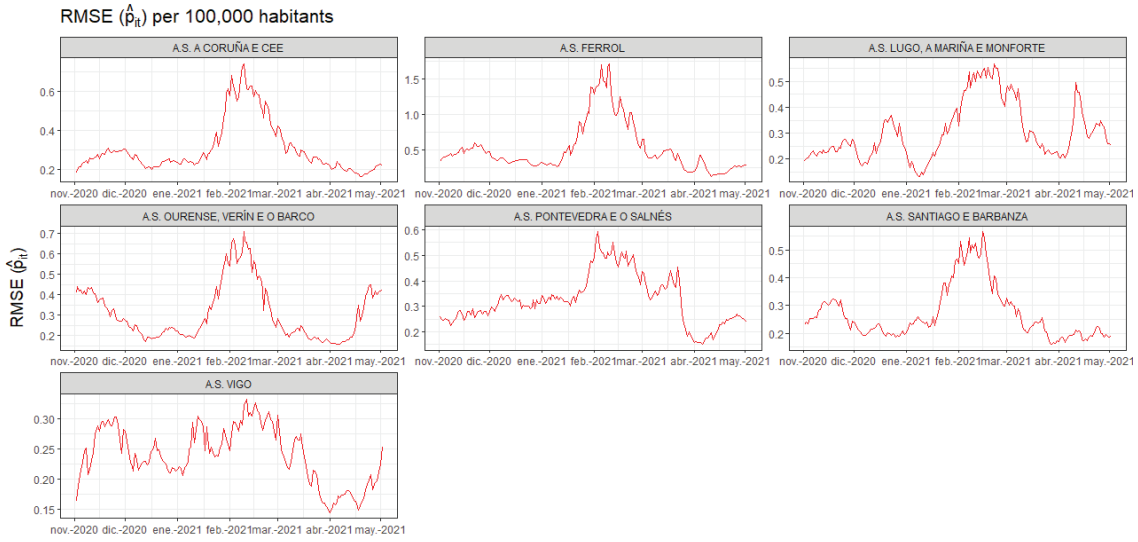


Figure 3: $RMSE$ of predicted ICU occupancy rate 100,000 inhabitants.

With the aim of further evaluating performance, we also proceeded to evaluate the $RRMSE$ as a percentage, calculated as the quotient of the $RMSE$ between the ICU occupancy rate estimated by the IN predictor for each area and day. Between February and March 2021, the $RRMSE$ takes values between 10 % and lower, being precisely the time range with lower pressure on care for all areas. With an average $RRMSE$ of 13.6620 %, values above 30 % are not reached in practically any case, except in the H.A. of Lugo, A Mariña e Monforte de Lemos, where values of 30% are reached occasionally in January and May 2021, and in the H.A. of Ferrol, in May 2021. In fact,

between April and May, all healthcare areas show an increase in $RRMSE$, which illustrates, again, that the predictor performs better in scenarios of high healthcare pressure than in low numbers of hospital admissions. Making a comparison between areas, both in terms of $RMSE$ and its relative measure, $RRMSE$, the error is greater in areas such as the H.A. of Ferrol, the H.A. of Pontevedra e o Salnés and the H.A. of Ourense, Verín e O Barco de Valdeorras (areas with the lowest number of inhabitants) compared to areas such as the H.A of A Coruña e Cee and the H.A. of Vigo, where greater precision is recorded (the two areas with the highest number of inhabitants).

5. INTENSIVE CARE UNIT OCCUPANCY FORECAST

In view of good estimation properties, it may be of interest to define what are known as out-of-sample predictors, consisting of predicting the values of the dependent variable y_{it} for $i \in \mathbb{I} = \{1, \ldots, I\}$ and $t \in \mathbb{T} = \{T+1, \ldots, t_o\}$, where t_o is the target time of the prediction of the out-of-sample prediction. Following the basic example of this theoretical explanation, it may be of interest to set as prediction targets time horizons of 7, 5 and 3 days, as these are the time windows commonly used to help better plan hospital resources or to implement non-pharmacological control measures (García-Vicuña et al., 2020). However, following the definition of the predictors, in order to make out-of-sample predictions it is necessary to establish the set $\boldsymbol{x}_{it}$ of explanatory variables in the prediction period, as well as to define the corresponding random intercepts, $u_{iw(t)}$.

Starting with the explanatory variables, different possibilities can be considered, but, in the context of application to real data, we resort to setting the values of the predictor variables t days before the prediction objective date t_o. As for the random effects at the intercept level, $\boldsymbol{u}_{iw(t)}$, there are two possibilities. Either ignore them, or if it is valuable to be kept, impute them. Due to their nature and importance, we evaluate several alternatives under simulation, such as imputation with the last, last two, with the mean of the last two, or with more elaborate methodologies such as imputation by exponential smoothing or by ARIMA models, by means of automatic algorithms implemented in the R forecast package (Hyndman et al., 2008). According to the results obtained with the bootstrap simulation we perform the imputation through the exponential smoothing.

It should be noted that this type of prediction has been previously performed in SAE in the context of forest fire prediction (Boubeta et al., 2019), but to our current knowledge, this is the first time it is defined for mixed models with random slopes and random effects imputation in this way. Thus, the forecasting algorithm could be as follows:

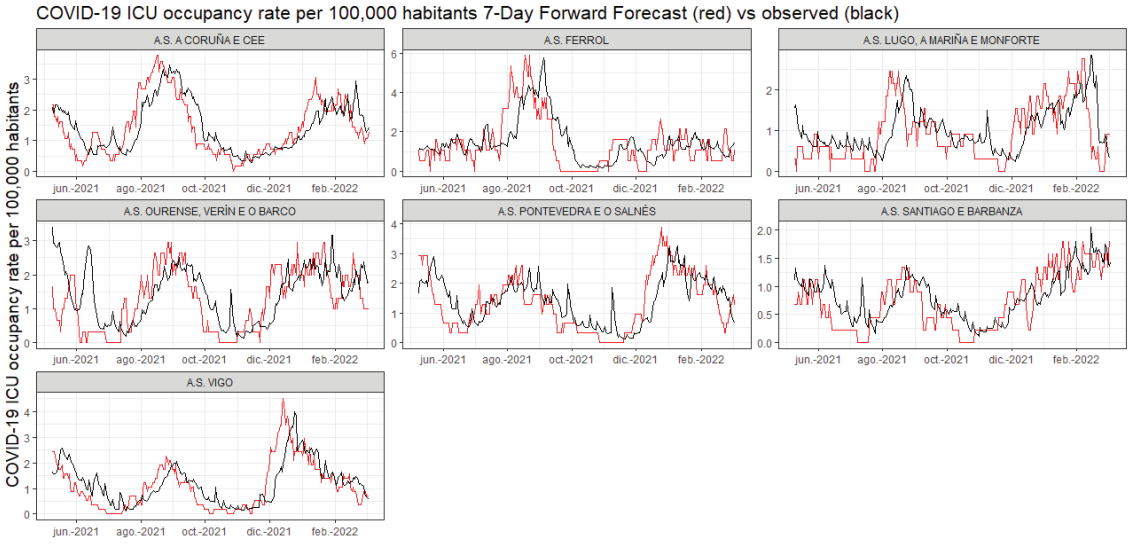
1. Fit the model to the sample and calculate the estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}, \hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\rho}}')'$, and the mode predictors $\hat{\boldsymbol{u}}_{iw(t)}, \hat{\boldsymbol{v}}_i$ for $i \in \mathbb{I} = \{1, \ldots, I\}$ and $t \in \mathbb{T} = \{t, \ldots, T\}$.

2. Set the target forecast threshold t_o, get the values of the auxiliary variables x_{it} for the t_o days prior to the origin of the prediciton, and impute the values of the random intercept, $\hat{\tilde{\boldsymbol{u}}}_{iw(t)}$ for $i \in \mathbb{I} = \{1, \ldots, I\}$ and $t \in \mathbb{T} = \{T+1, \ldots, t_o\}$.

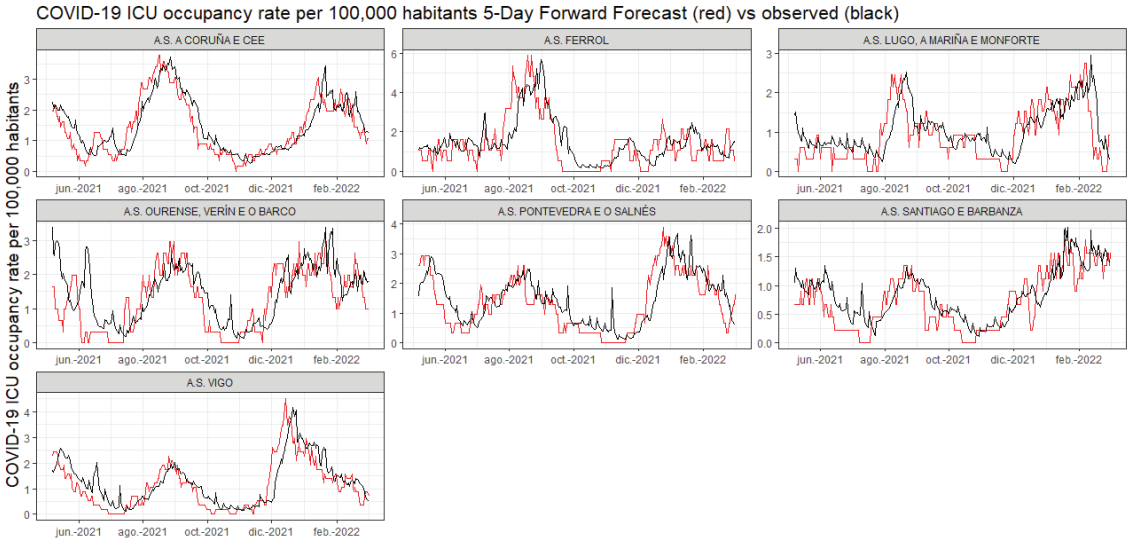
3. Calculate
$\hat{\tilde{p}}^{\text{jn}}_{it} = \exp\{\boldsymbol{x}_{it-t_o}\hat{\boldsymbol{\beta}} + \hat{\sigma}\hat{\tilde{\boldsymbol{u}}}_{iw(t)} + \sum_{\ell=1}^{p} \hat{\phi}_\ell \hat{v}_{\ell,i} x_{\ell,it-t_o}\}$ for $i \in \mathbb{I} = \{1, \ldots, I\}$ and $t \in \mathbb{T} = \{T+1, \ldots, t_o\}$.

Preliminary results, show that although the trend of the observed data is reproduced, there is a time lag in the prediction, which we consider to be the result of the use of lagged values in the auxiliary variables. This effect is alleviated as the forecast horizon is shortened to five, and especially three days, as visualised in Figures 4a, 4b, 4c. A worse performance is also discernible from January 2022 onwards. Although an in-depth analysis is required, it should be noted that as of 9 January 2022, SERGAS also counted home test results among the positives (Lipido, 2023). This affects the number of people discharged, a variable used in the model.

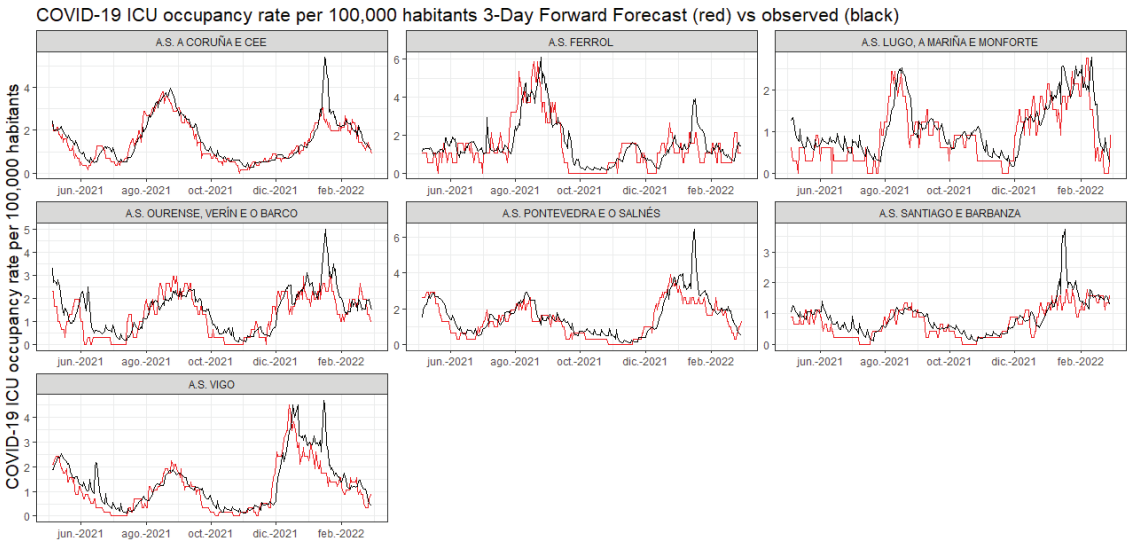
To further assess the accuracy achieved, we estimate the $RMSE$ using a parametric bootstrap with 500 bootstrap replicates, as this is the best measure of error that can be provided to the user in real time, as the number of inpatients to be registered on the target day is not available at the time of prediction. For the worst case, 7-day forecast, with an average $RMSE$ of 1.3332 hospitalised in ICU per 100,000 inhabitants, with a minimum of 0.2897 and a maximum of 6.9875 (reached in the the H.A. of Pontevedra e o Salnés in December 2021), the performance is optimal



(a) Observed (black line) vs 7-Day forward forecast (red line) ICU occupancy rate per 100,000 inhabitants.



(b) Observed (black line) vs 5-Day forward forecast (red line) ICU occupancy rate per 100,000 inhabitants.



(c) Observed (black line) vs 3-Day forward forecast (red line) ICU occupancy rate per 100,000 inhabitants.

Figure 4: Comparison of ICU occupancy rate forecasts.

in both high pressure and low pressure care scenarios, which motivates us to conduct future studies in which we improve the random effect imputation technique and auxiliary variable values.

As an example, Figures 5, 6 and 7 are provided, which map, respectively, the observed COVID-19 ICU occupancy rate per 100,000 habitants, the predicted rate seven days ahead and the bootstrap $RMSE$ associated with the prediction, illustrating the good extension to all types of settings, with a tendency to overestimate, which is preferable to underestimate in this context.

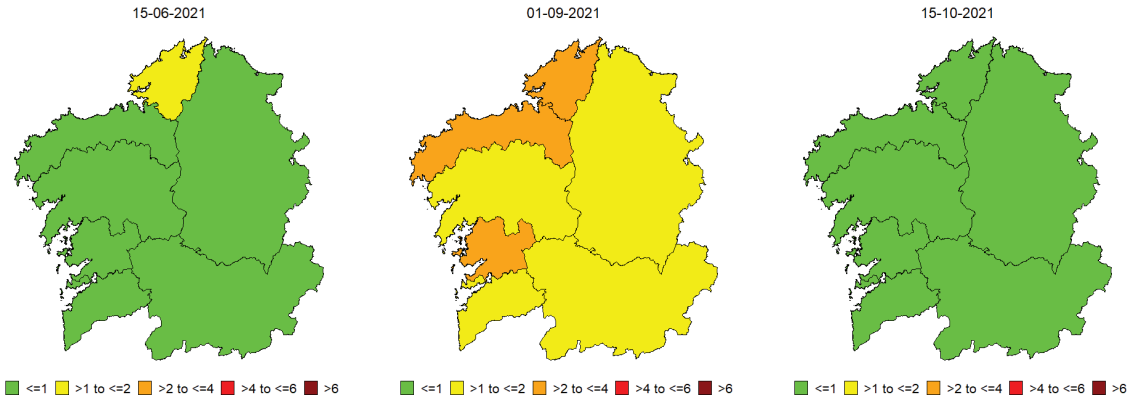


Figure 5: Observed COVID-19 ICU occupancy rate per 100,000 habitants.

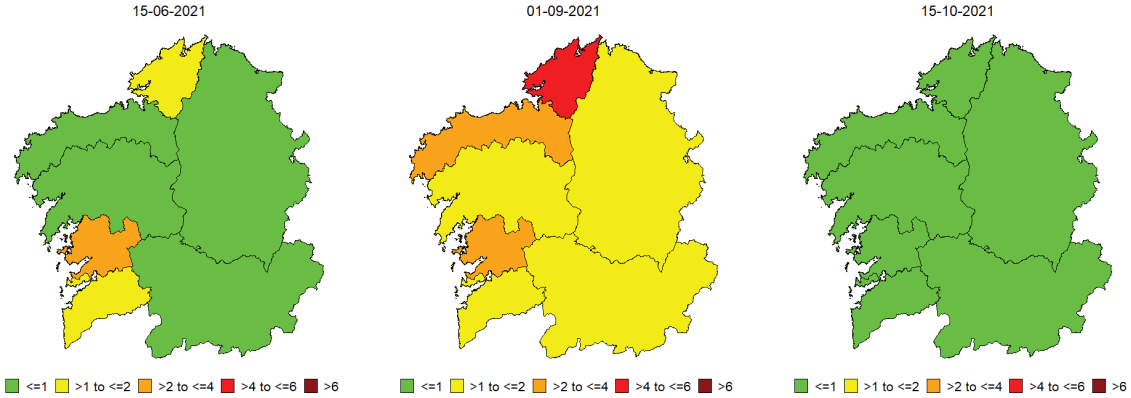


Figure 6: Predicted COVID-19 ICU occupancy rate per 100,000 habitants.

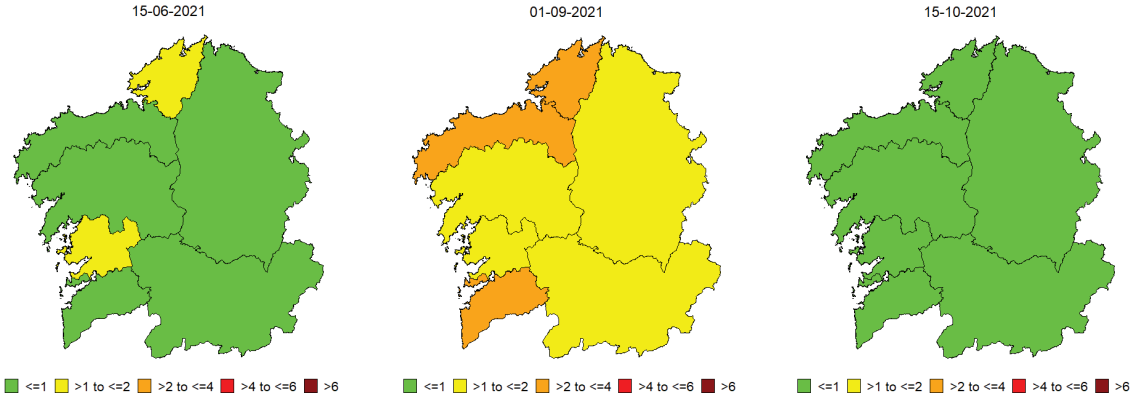


Figure 7: $RMSE$ of the predicted ICU occupancy rate per 100,000 habitants estimated by bootstrap.

It should be noted that the selection of the map scale refers to the risk levels relative to the ICU occupancy rate per 100,000 inhabitants specified in the different state strategies published

by the Ministry of Health throughout the months of the pandemic. However, as highlighted by the health authorities, in order to establish whether confinement or restriction measures should be applied, the risk assessment must also consider the specific characteristics of the territorial unit being assessed, putting it in context with other indicators (Ministerio de Sanidad, 2022).

6. CONCLUSIONS

In this paper, we apply, for the first time, a Poisson-type area model with random regression coefficients to model the daily overload in intensive care units as a consequence of COVID-19 disease in each of the seven health areas of the Autonomous Community of Galicia, assessing both modelling and forecast capacity.

In terms of accuracy in modelling the occupancy of intensive care units, the predictors show excellent performance, both in simulations and in application to real data. Thus, in the period between 2nd November 2020 and 2nd May 2021 there is a heterogeneity in the impact of COVID-19 depending on each area, with the Health Area of A Coruña e Cee and the Health Area of Ferrol standing out for their higher rates. Furthermore, although the precision of the estimates is better in scenarios of high care pressure, it is worth noting the good performance in terms of relative root mean squared error in all types of scenarios.

Regarding forecasting the future, we started the evaluation of the performance of the predictors in the elaboration of future predictions in time horizons of 3, 5 and 7 days, for each of the health areas in the period between 3rd May 2021 and 6th March 2022. To our current knowledge, this is the first time that is maked for generalized linear mixed models with random slopes and random effects imputation and the results, in terms of root mean squared error obtained by parametric bootstrap, are encouraging in both low and high care pressure scenarios. Therefore, we are currently developing a Shiny application to obtain these predictions at user level to facilitate resource planning.

In the future, we plan to evolve this approach to perform modelling and prediction at a higher level of disaggregation, such as health district and even hospital, as well as to make improvements in the prediction algorithm by exploring alternatives for the imputation of auxiliary variables and unknown random effects. Thus, we believe that these approximations will also improve the 5-day and 7-day forward forecasts. In addition, we will continue to investigate optimisation methods or other alternative approaches to EBP and BP, since, despite the high computational cost, their performance could be within the quality standards needed for this problem.

ACKNOWLEDGMENTS

This research is part of the grant PID2020-113578RB-I00, funded by `MCIN/AEI/10.13039/501100011033/`. It has also been supported by the Spanish grant PID2022-136878NB-I00, the Valencian grant Prometeo/2021/063, by the Xunta de Galicia (Competitive Reference ED431C-2020/14) and by CITIC that is supported by Xunta de Galicia, collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Sistema Universitario de Galicia (CIGUS). The first author was also sponsored by the Spanish Grant for Predoctoral Research Trainees RD 103/2019 being this work part of grant PRE2021-100857, funded by MCIN/AEI/10.13039/501100011033/ and ESF+.

REFERENCES

Boubeta, M., Lombardía, M. J., Marey-Pérez, M., and Morales, D. (2019) Poisson mixed models for predicting number of fires. International journal of wildland fire, 28(3), 237–253.

Centro Nacional de Epidemiología (CNE). (2022) Factores COVID-19. ISCIII. Available at: `https://coviddifusion.isciii.es/fdd/`. (Accessed: July 2023).

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

Fay, R. E., and Herriot, R. A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269–277.

García-Vicuña, D., Esparza, L., and Mallor, F. (2022) Hospital preparedness during epidemics using simulation: the case of COVID-19. Central European Journal of Operations Research, 30(1), 213–249.

García-Vicuña, D., López-Cheda, A., Jácome, M. A., and Mallor, F. (2023) Estimation of patient flow in hospitals using up-to-date data. Application to bed demand prediction during pandemic waves. PLoS One, 18(2), e0282331.

García-Vicuña, D., Mallor, F., and Esparza, L. (2020, December) Planning ward and intensive care unit beds for COVID-19 patients using a discrete event simulation model. In 2020 Winter Simulation Conference (WSC) (pp. 759–770). IEEE.

Hyndman, R. J., and Khandakar, Y. (2008) Automatic time series forecasting: the forecast package for R. Journal of statistical software, 27, 1–22.

Ibáñez, M. V., Martínez-García, M., and Simó, A. (2021) A review of spatiotemporal models for count data in R packages. A case study of COVID-19 data. Mathematics, 9(13), 1–23.

INE (2021) Population and Housing Censuses 2021. Available at: `https://www.ine.es/dynt3/inebase/en/index.htm?padre=8952`. (Accessed: July 2023).

Lipido. (2023) Datos COVID-19 Galicia. Available at GitHub: `https://github.com/lipido/galicia-covid19`. (Accessed: July 2023).

Lombardía, M.J., López-Vizcaíno, E., and Rueda, C. (2017) Mixed generalized Akaike information criterion for small area models. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(4), 1229-1252.

López-Cheda, A., Jácome, M. A., Cao, R., and De Salazar, P. M. (2021) Estimating lengths-of-stay of hospitalised COVID-19 patients using a non-parametric model: a case study in Galicia (Spain). Epidemiology & Infection, 149, E102.

Martínez-Beneito, M.A., Mateu, J., and Botella-Rocamora, P. (2022) Spatio-temporal small area surveillance of the COVID-19 pandemic. Spatial Statistics, 49, 100551.

Ministerio de Sanidad. (2022) Estrategia de vigilancia y control frente a COVID-19 tras la fase aguda de la pandemia. (Revisado a 8 de noviembre de 2022). Available at: `https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Nueva_estrategia_vigilancia_y_control.pdf`. (Accessed: July 2023).

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

Pfeffermann, D. (2013) New important developments in small area estimation. Statistical Science, 28(1), 40–68.

Rao, J.N.K., and Molina, I. (2015) Small area estimation. 2nd ed. Wiley, Hoboken.

Safari, W. C., López-de-Ullibarri, I., and Jácome, M. A. (2022) Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed. Statistical Methods in Medical Research, 31(11), 2164–2188.

Servizo Galego de Saúde (SERGAS). (2023) Datos Coronavirus. Available at: `https://coronavirus.sergas.es/datos/#/gl-ES/galicia`. (Accessed: July 2023).

Trigo-Tasende, N., Vallejo, J. A., Rumbo-Feal, S., Conde-Pérez, K., Vaamonde, M., López-Oriona, Á., Barbetio, I., Nasser-Ali, M., Reif, R., Rodiño-Janeiro, B.K., Fernández-Álvarez, E., Iglesias-Corrás, I., Freire, B., Tarrío-Saavedra, J., Tomás, L., Gallego-García, L., Posada, D., Bou, G., López-de-Ullibarri, I., Cao, R., Ladra, S., and Poza, M. (2023) Wastewater early warning system for SARS-CoV-2 outbreaks and variants in a Coruña, Spain. Environmental Science and Pollution Research, 1–20.

Vaida, F., and Blanchard, S. (2005) Conditional Akaike information for mixed-effect models. Biometrika, 92(2), 351–370.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

ALGORITMOS DE APRENDIZAJE BASADOS EN ÁRBOLES DE EXPANSIÓN MÍNIMA

Iria Rodríguez Acevedo[1], Julio González Díaz[1] y Beatriz Pateiro López[1].

[1] Departamento de Estadística, Análisis Matemático y Optimización, grupo de investigación MODESTYA, Universidad de Santiago de Compostela.

RESUMEN

El problema de clasificar una observación en función de sus características observables en cierta clase o categoría es una tarea fundamental del análisis de datos. Su gran cantidad de aplicaciones hace que sea objeto de estudio mediante el desarrollo de diferentes técnicas de clasificación. El objetivo de este trabajo es presentar una nueva técnica de clasificación basada en árboles de expansión mínima, analizando su comportamiento mediante un exhaustivo estudio computacional.

Palabras y frases clave: k-vecinos más próximos, regla de clasificación, consistencia universal, regla de Bayes, árboles de expansión mínima, estudio computacional.

1. INTRODUCCIÓN

Este trabajo se encuadra en la intersección de dos ramas de las matemáticas que, si bien son distintas, poseen relación: la estadística y la investigación operativa. Más concretamente, el objetivo del trabajo será presentar una nueva técnica de clasificación basada en árboles de expansión mínima.

Por una parte, el origen de los problemas de clasificación se remonta hasta épocas muy lejanas cuando los humanos comenzaban a clasificar objetos según el metal de fabricación, categorizar a los seres vivos (taxonomía) o mismo agrupar a los alimentos según su función o propiedades. A pesar de ello, desde el punto de vista de la estadística y del análisis de datos el estudio de los problemas de clasificación es más contemporáneo, impulsado sobre todo en los últimos tiempos por los avances tecnológicos y computacionales que permitieron desarrollar nuevas técnicas que hasta el momento eran inviables.

En los años 30 surgió una de las primeras reglas de clasificación de la mano del estadístico y biólogo Ronald Aylmer Fisher: el análisis lineal discriminante (Fisher 1936). Dicha técnica tiene como objetivo separar las clases mediante hiperplanos. Avanzando un poco en el tiempo, ya en los años 50, surgen otras técnicas de clasificación como k-vecinos más próximos (Fix y Hodges 1951) o el algoritmo Perceptrón (Rosenblatt 1958). Este último es el impulsor de lo que hoy en día conocemos como redes neuronales. En los años 70 y 80 se desarrollaron otras técnicas como los árboles de decisión, siendo pionero en este campo John Ross Quinlan con la invención de algoritmos de clasificación como ID3 (Quinlan 1979), que derivaron en la creación de los árboles de decisión (Quinlan 1986). En la década de los 90 surge la técnica de *support vector machines* (SVM) en los trabajos Boser et al. (1992) y Cortes y Vapnik (1995), impulsando así la clasificación binaria. A finales de la década de los 90 se empezaron a desarrollar técnicas de ensamblado, de forma que no se tenga en cuenta la clasificación devuelta por un solo clasificador. Uno de los clasificadores más conocidos que surgieron de aplicar esta idea son los bosques aleatorios o *random forests* (Breiman 2001).

En las últimas décadas, con el desarrollo de internet y de la computación, se ha experimentado un notable crecimiento en el estudio del conocido como *machine learning* o aprendizaje automático. Esta es una rama de la inteligencia artificial que se ocupa del desarrollo de algoritmos de aprendizaje, dentro de los cuales se encuentra la clasificación. Se impulsa entonces el estudio de nuevos métodos de clasificación, como aquellos basados en redes neuronales.

La gran cantidad de clasificadores que se han ido desarrollando a lo largo de la historia se debe en gran medida a la amplia variedad de aplicaciones que podemos encontrar en múltiples contextos, como pueden ser la detección de spam de un correo electrónico, en sistemas de detección de fraudes o en finanzas.

Por otra parte, encontrar el árbol de expansión mínima de un grafo es uno de los problemas combinatorios más conocidos en el campo de la optimización. Un árbol de expansión mínima es un subconjunto de aristas de un grafo no dirigido, conexo y ponderado que conecta todos los vértices entre sí, sin ningún ciclo y con el mínimo peso total posible de las aristas. Este concepto surgió a raíz del desarrollo de la teoría de grafos a principios del siglo veinte, durante el cual fueron surgiendo diferentes algoritmos para su obtención. El primero de ellos fue introducido por el matemático checo Otakar Boruvka en 1926 (Boruvka 1926). En 1930 el matemático de procedencia también checa Vojtech Jarník propuso otro algoritmo que posteriormente Robert Clay Prim modificó dando lugar a uno de los algoritmos para la obtención de árboles de expansión mínima más conocidos: el algoritmo de Prim (Prim 1957). Casi de forma paralela, en 1956, Joseph B. Kruskal presentó otro algoritmo también muy extendido: el algoritmo de Kruskal (Kruskal 1956).

Todos estos algoritmos son voraces, es decir, son algoritmos que realizan elecciones localmente óptimas en cada etapa y dan como resultado una solución globalmente óptima. Su fácil implementación y rapidez hace que surjan diversas aplicaciones: en redes de comunicación, en rutas de transporte o para la creación de mapas. Hoy en día se continúan estudiando las propiedades de los árboles de expansión mínima, investigando nuevas aplicaciones que se valgan de ellas. Una de estas nuevas aplicaciones es precisamente la técnica de clasificación que presentaremos. La estructura de los árboles de mínimo coste nos permitirá captar la configuración de cada clase.

2. REGLA BASADA EN ÁRBOLES DE EXPANSIÓN MÍNIMA

Es conocido que existen muchas aplicaciones de los árboles de expansión mínima. Una de las principales razones es su fácil construcción gracias a algoritmos voraces como el de Prim. Sin embargo, no hemos encontrado en la literatura referencias que utilicen árboles de expansión mínima para definir reglas de clasificación. Sí que hemos identificado trabajos en los que se utilizan árboles de expansión mínima en aprendizaje no supervisado, véase Zahn (1971). Lo que se hace en ese caso es calcular el árbol de expansión mínima de las observaciones y después, si se quieren formar k clústers, se suprimen las $k-1$ aristas con coste más alto del mismo. Las k componentes conectadas resultantes constituyen los clústers. Nos parece natural pensar que también se podría explotar el uso de árboles de expansión mínima en el contexto de aprendizaje supervisado. En este sentido, la regla de clasificación binaria que presentaremos en este trabajo constituye una contribución metodológica novedosa de este trabajo. Dado un conjunto de entrenamiento, la idea del método se basa en la construcción de árboles de expansión mínima para cada una de las clases, asignando una nueva observación a aquella clase cuyo árbol de expansión se vea menos afectado al introducir la nueva observación en ella. La forma de medir esta conformidad se llevará a cabo mediante lo que se denominará separación de un grafo. Dicha separación se define como el cociente entre el coste de los árboles de expansión mínima del grafo y el número de observaciones. Así, la conformidad de cada clase será el cociente entre la separación del grafo y la separación del mismo una vez añadida la nueva observación. A mayor conformidad, más afín será, lo que determinará la clasificación. Dicho procedimiento es fácilmente extendible a más de dos clases.

En primer lugar, se dará la definición formal de la regla. A continuación se presentará una primera mejora del método, haciéndolo más robusto frente a observaciones atípicas y que, al mismo tiempo, lo hace mucho más eficiente desde el punto de vista computacional. Por último, se muestra un estudio numérico de la regla junto con las conclusiones obtenidas.

2.1. Definición de la regla

Considérese el vector aleatorio (X, Y) que toma valores en $\mathbb{R}^d \times \{0, 1\}$ y supongamos que disponemos de $(X_1, Y_1), \ldots, (X_n, Y_n)$, una secuencia de pares aleatorios independientes e idénticamente distribuidos a (X, Y). Se denotará por n_p al número de observaciones de la muestra tales que $Y_i = 1$ (observaciones positivas) y por n_n al número de observaciones tales que $Y_i = 0$ (observaciones negativas).

El primer paso de la regla consiste en normalizar todas las variables explicativas (en total d por ser X de esta dimensión), restándoles su media (muestral) y dividiendo por la desviación típica (muestral). De esta forma tendremos a todas ellas en la misma escala.

El segundo paso es definir los grafos correspondientes a cada clase y calcular árboles de expansión mínima para cada uno:

- Por GP denotaremos al grafo no dirigido completo cuyos nodos vienen dados por las observaciones de la clase 1. Los costes de las aristas vienen dados por las distancias euclideas entre ellas. Denotaremos por c_p al coste de cualquier árbol de expansión mínima de GP (todos ellos tienen el mismo coste).

- De forma análoga, por GN denotaremos al grafo no dirigido completo cuyos nodos vienen dados por las observaciones de la clase 0. Los costes de las aristas vienen dados por las distancias euclideas entre ellas. Denotaremos por c_n al coste de cualquier árbol de expansión mínima de GN.

El siguiente paso es definir una variable que mida la separación entre observaciones en cada grafo. Para el primero de ellos se define $SEP(GP) = \frac{c_p}{n_p}$. Para el segundo grafo por su parte $SEP(GN) = \frac{c_n}{n_n}$. Es decir, se divide el coste del árbol entre el número de observaciones. De esta forma, cuanto menor sea el valor de SEP más semejantes serán entre sí las observaciones de la clase en cuestión.

Dada una observación x, para clasificarla en la clase de las observaciones positivas o en la de las negativas se procede como sigue:

- Se definen los grafos GP^x y GN^x como los grafos resultantes tras añadir x tanto a la clase de las observaciones positivas como a la de las negativas.

- Se calculan $SEP(GP^x) = \frac{c_p^x}{n_p+1}$ y $SEP(GN^x) = \frac{c_n^x}{n_n+1}$, siendo c_p^x y c_n^x los costes de cualquier árbol de expansión mínima de GP^x y GN^x, respectivamente.

- Como último paso se define una variable que mide la mejora o empeoramiento de la separación en cada uno de los grafos al añadir la nueva observación:

$$CONF_P^x = \frac{SEP(GP)}{SEP(GP^x)} \quad \text{y} \quad CONF_N^x = \frac{SEP(GN)}{SEP(GN^x)}.$$

  Estas variables se denominarán *conformidad en la clase 1* y *conformidad en la clase 0*, respectivamente. A mayor valor, mayor conformidad, es decir, menor separación entre los datos en GP^x/GN^x relativa a la separación en GP/GN.

Se define entonces la regla de clasificación basada en árboles de expansión mínima, de forma abreviada *MST-Class* (*Minimum Spanning Tree Classifier*), como:

$$g_n(x) = \begin{cases} 1 & \text{si } CONF_P^x > CONF_N^x, \\ 0 & \text{si } CONF_P^x < CONF_N^x. \end{cases}$$

En el caso en el que $CONF_P^x = CONF_N^x$ se aleatoriza la clasificación.

2.2. Método robusto

Una primera mejora que se ha incluido en el método viene motivada por intentar solventar un posible problema de robustez de la regla frente a observaciones atípicas. Se pueden consultar resultados más detallados sobre este aspecto en el estudio de simulación que se presenta más adelante, pero a efectos ilustrativos, presentamos a continuación un pequeño ejemplo en el que se hace evidente el problema. Supongamos la siguiente situación en la que la observación (X_4, Y_4) está mal clasificada en la clase 1 en vez de en la 0:

$$(X_1, Y_1) = ((0,0),0), (X_2, Y_2) = ((1,0),0), (X_3, Y_3) = ((0,1),0), (X_4, Y_4) = ((0{,}5, 0{,}5),1),$$
$$(X_5, Y_5) = ((5,5),1), (X_6, Y_6) = ((6,5),1), (X_7, Y_7) = ((1,1),0).$$

En primer lugar, reescalemos todas las variables explicativas, restándoles su media muestral y dividiendo por la desviación típica muestral. El conjunto de datos resultante se muestra en la Figura 1. Puede apreciarse que la observación (X_4, Y_4), al estar etiquetada con $Y_4 = 1$, se encuentra en una zona en la que las observaciones pertenecen a la otra clase.

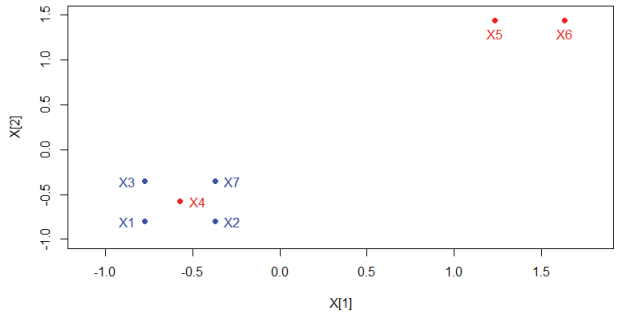


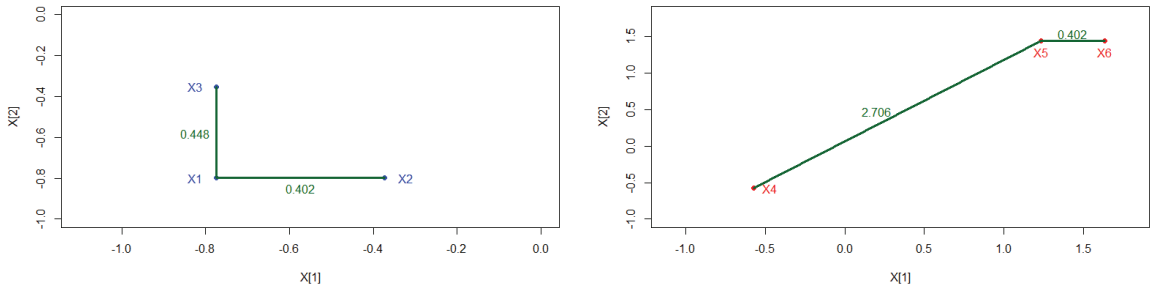
Figura 1: Representación de las observaciones: en rojo las de la clase 1 y en azul las de la 0.

Consideremos ahora como muestra de entrenamiento el conjunto $(X_i, Y_i), i = 1, \ldots, 6$ y veamos qué clasificación resulta para la observación (X_7, Y_7) según la regla *MST-Class*. Para calcular los árboles de mínimo coste en primer lugar se calculan las distancias entre las X_i:

$$d((X_1, X_2)) = 0{,}4015326, d((X_1, X_3)) = 0{,}4477469, d((X_2, X_3)) = 0{,}6014198$$
$$d((X_4, X_5)) = 2{,}7063889, d((X_4, X_6)) = 2{,}9894523, d((X_5, X_6)) = 0{,}4015326.$$

Los árboles de mínimo coste se calculan a simple vista en este ejemplo pequeño, resultando los árboles representados en la Figura 2a y en la Figura 2b .Así, $c_n = 0{,}4015326 + 0{,}4477469 = 0{,}8492795$ y $c_p = 2{,}7063889 + 0{,}4015326 = 3{,}107921$. Por tanto,

$$SEP(GP) = \frac{c_p}{n_p} = \frac{3{,}107921}{3} = 1{,}035974, \quad SEP(GN) = \frac{c_n}{n_n} = \frac{0{,}8492795}{3} = 0{,}2830932.$$



(a) Representación del árbol de expansión mínima para la clase 0, en verde.

(b) Representación del árbol de expansión mínima para la clase 1, en verde.

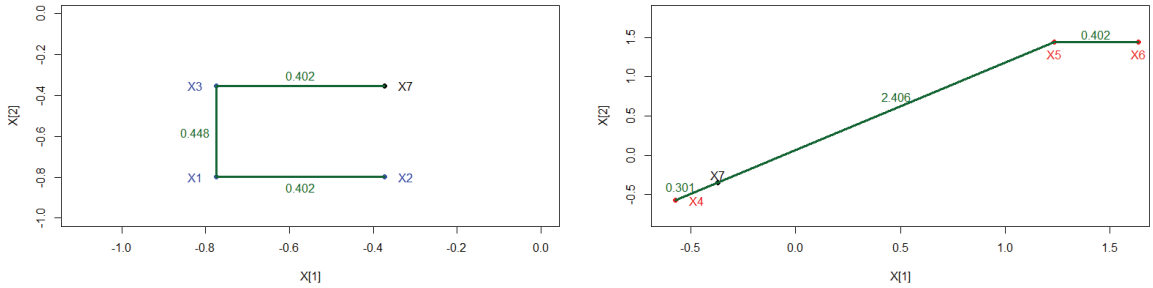
Figura 2: Representaciones de los árboles de expansión mínima.

Calculemos ahora los grafos GP^{X_7} y GN^{X_7}. De nuevo, se calculan las distancias entre las X_i:

$$d((X_1, X_2)) = 0{,}4015326, d((X_1, X_3)) = 0{,}4477469, d((X_1, X_7)) = 0{,}6014198,$$
$$d((X_2, X_3)) = 0{,}6014198, d((X_2, X_7)) = 0{,}4477469, d((X_3, X_7)) = 0{,}4015326$$
$$d((X_4, X_5)) = 2{,}7063889, d((X_4, X_6)) = 2{,}9894523, d((X_4, X_7)) = 0{,}3007099,$$
$$d((X_5, X_6)) = 0{,}4015326, d((X_5, X_7)) = 2{,}4056791, d((X_6, X_7)) = 2{,}6904177.$$

De nuevo, podemos calcular los árboles de mínimo coste a simple vista, resultando los árboles representados en la Figura 3a y en la Figura 3b. Así, $c_n^{X_7} = 0{,}4015326 + 0{,}4015326 + 0{,}4477469 = = 1{,}250812$ y $c_p^{X_7} = 0{,}3007099 + 2{,}4056791 + 0{,}4015326 = 3{,}107922$. Por consiguiente,

$$SEP(GP^{X_7}) = \frac{c_p}{n_p} = \frac{3{,}107921}{4} = 0{,}7769804, \quad SEP(GN^{X_7}) = \frac{c_n}{n_n} = \frac{1{,}250812}{4} = 0{,}312703.$$



(a) Representación (en verde) del árbol de expansión mínima para la clase 0 añadiendo X_7.

(b) Representación (en verde) del árbol de expansión mínima para la clase 1 añadiendo X_7.

Figura 3: Representaciones de los árboles de expansión mínima.

Como último paso se calculan las variables de conformidad:

$$CONF_P^{X_7} = \frac{SEP(GP)}{SEP(GP^{X_7})} = \frac{1{,}035974}{0{,}7769804} = 1{,}333334$$
$$CONF_N^{X_7} = \frac{SEP(GN)}{SEP(GN^{x})} = \frac{0{,}2830932}{0{,}312703} = 0{,}9053102.$$

Por tanto, la observación X_7 se clasificaría en la clase $Y = 1$, cuando realmente $Y_7 = 0$ y todo como consecuencia de tener X_4 mal etiquetado. Para mitigar este problema y además hacer que la regla de clasificación use más información que simplemente la información sobre el grafo completo se procede como se describe a continuación.

Dada una observación x:

i) Se definen aleatoriamente l submuestras de la muestra de entrenamiento. Para que en dichas submuestras las clases estén balanceadas se procede como sigue:

- Se define $n \coloneqq \text{mín}\{n_p, n_n\}$.
- Se escogen aleatoriamente $\sqrt{n}$ observaciones positivas y $\sqrt{n}$ observaciones negativas.

Esta elección de tamaño para las submuestras es de especial importancia dado el orden de complejidad de los algoritmos para el problema del árbol de expansión mínima. Dicha complejidad depende de ciertos detalles de la implementación en los que no vamos a entrar. Sin embargo, en ningún caso es peor que $O(n^2)$, que es el valor que tomaremos de referencia.

Por tanto, calcular los árboles de expansión mínima necesarios tiene una complejidad de $O(n_p^2)$ para la clase de las observaciones positivas, mientras que para las negativas es de $O(n_n^2)$. Por tanto, si pasamos a calcularlos con l submuestras de tamaño $\sqrt{n}$ en cada clase, la complejidad computacional para cada submuestra pasa a ser de $O((\sqrt{n})^2) = O(n)$ en ambas clases. De esta forma, el tiempo de cálculo pasa de ser cuadrático en el número de observaciones a ser lineal en el número de observaciones (que habría que multiplicar por l, que para conjuntos de datos grandes será mucho más pequeño que n_p y n_n).

Esta mejora computacional es especialmente relevante en contextos *bigdata*, en los que pasar de una complejidad cuadrática a una lineal tiene un gran impacto.

ii) Para cada una de las l submuestras se calculan los valores $CONF_P^x$ y $CONF_N^x$.

iii) Como siguiente paso se calcula la media ponderada de estos valores.

iv) Finalmente se realiza la clasificación en base a esta medida agregada de conformidad.

A este clasificador lo denotaremos de forma abreviada por *MST-RClass* (*Minimum Spanning Tree Robust Classifier*).

Volviendo al ejemplo, esto resultaría en que muchas de las submuestras tomadas dejarían fuera a X_4. Por tanto, en esos casos la conformidad de X_7 con las observaciones azules ($Y_i = 0$) será mucho mejor que con las rojas ($Y_i = 1$). Como consecuencia, al promediar los valores de conformidad devueltos por todas las submuestras seguramente acabaríamos realizando la clasificación de forma correcta.

3. ESTUDIO COMPUTACIONAL

Para llevar a cabo el estudio computacional se ha tenido que implementar de cero tanto la regla de clasificación base como la robusta. Para ello se ha elaborado el código necesario empleando el software R (R Core Team 2023). Cabe mencionar que la librería empleada para calcular el árbol de expansión mínima es la librería *igraph* (Csárdi et al. 2023), la cual utiliza en su función *mst* el algoritmo de Prim. Asimismo, se ha elaborado el código necesario para generar todas las gráficas presentes en el trabajo empleando la función *plot3d* de la librería *rgl* (Adler et al. 2023) y la función *ggplot* de la librería *ggplot2* (Chang et al. 2023). El cálculo de las medidas de rendimiento que emplearemos para comparar los distintos métodos ha sido realizado también empleando el software R. Por otra parte, todos los conjuntos de datos generados han sido almacenados en achivos *.csv*. Además, tanto en las ejecuciones realizadas como en el código generador de los diferentes conjuntos de datos se han establecido semillas empleando la función *set.seed* del paquete básico de R (R Core Team 2023), de forma que se puedan reproducir todas las ejecuciones de nuevo si fuese preciso.

Dado que el número de ejecuciones a realizar era elevado y el tiempo de las mismas era demasiado extenso como para llevarlas a cabo en un ordenador estándar, todas ellas se han realizado en el superordenador Finisterrae III, proporcionado por el Centro de Supercomputación de Galicia (CESGA). Salvo para cuando se ha querido comparar el rendimiento de las diferentes configuraciones del método no se han empleado nodos exclusivos, y el tiempo de ejecución máximo establecido así como la capacidad de memoria RAM del nodo se han adaptado a cada ejecución (conjuntos de datos con tamaños más pequeños requieren menos tiempo y capacidad).

3.1. Comparaciones entre distintas configuraciones de *MST-RClass* y *MST-Class*

El primer estudio que se ha realizado ha sido de tipo comparativo entre diferentes configuraciones de la versión robusta y la versión estándar de la regla. En particular, se han considerado 3 configuraciones distintas para *MST-RClass* variando el número de submuestras l (todas ellas con $\sqrt{n}$ observaciones positivas y $\sqrt{n}$ negativas, $n = \text{mín}\{n_n, n_p\}$): $l = 10$ (*MST-RClass_10*), $l = 50$ (*MST-RClass_50*) y $l = 100$ (*MST-RClass_100*).

Para testear el rendimiento de estas configuraciones hemos generado en una primera instancia conjuntos de datos con varias clases, donde la distribución subyacente de cada clase es una distribución normal multivariante preestablecida. En particular, se han generado 100 conjuntos de datos por cada una de las configuraciones. Se han escogido varios tamaños para el número de observaciones de cada clase en cada conjunto de datos generado: 300, 600, 1500 y 3000. Denotaremos por n_i al tamaño de la clase i. Además, las diferentes configuraciones surgen de ir realizando cambios progresivos sobre los parámetros de dos distribuciones normales que constituyen una estructura de referencia. Para la primera de las clases dichos parámetros son $\mu_1 = (1, 2)$, y

$$\Sigma_1 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix},$$

mientras que para la segunda clase $\mu_2 = (2, 3)$, y

$$\Sigma_2 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}.$$

De forma aleatoria, dentro de cada conjunto de datos se ha seleccionado una muestra de entrenamiento con una proporción del 70 % del total, dejando un 30 % como muestra test sobre la que se han evaluado los resultados de clasificación. Al final del clasificado se calcula la proporción de bien clasificados para ese conjunto de datos. Después de repetir el proceso para los 100 conjuntos de datos de cada configuración se hace un resumen sobre todas las proporciones obtenidas.

Todos los resultados de las ejecuciones se han recogido en forma de tablas en el Capítulo 1, Sección 1,1, del archivo del siguiente repositorio: `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`.

Además del resumen sobre las proporciones de bien clasificados se ha añadido una fila más a las tablas titulada *Elapsed*. Esta fila recoge el tiempo medio, en segundos, de las 100 ejecuciones de cada configuración de datos. Para medir dicho tiempo se han solicitado nodos exclusivos en el CESGA. Además, se ha empleado la función *proc.time* incluida en el paquete básico de R, que devuelve un vector de dimensión 5 con diferentes tiempos medidos como el de usuario, el de sistema y el real transcurrido. El tiempo de usuario es el tiempo de CPU relacionado con la ejecución del código, mientras que el tiempo de sistema se relaciona con procesos como abrir o cerrar archivos. Para nuestro estudio hemos empleado el tiempo real transcurrido o *real elapsed time*.

Con respecto a las proporciones de bien clasificados podemos concluir que el método robusto proporciona mucho mejores resultados que la regla original, aún cuando no se está en presencia de datos atípicos o *outliers*. Esto constituye una gran ventaja, ya que no solo devuelve mejores resultados sino que además será más eficiente computacionalmente, tal y como veremos. Además, a mayor número de submuestras mejores resultados se obtuvieron. En particular, se observa un mayor impacto al considerar $l = 50$ en vez de $l = 10$ que el observado al considerar 100 submuestras en vez de 50.

Con respecto a los tiempos de ejecución se ha observado lo siguiente:

- Cuando el tamaño del conjunto de datos es pequeño ($n_1 = n_2 = 300$) tan solo se aprecia una reducción del tiempo transcurrido con respecto a la regla original cuando el número de submuestras es pequeño, es decir, $l = 10$. Sin embargo, cuando $l = 50$ o $l = 100$ el tiempo transcurrido se eleva, incluso triplicando el tiempo de la regla base. Esto se relaciona con la complejidad computacional del método robusto que hemos comentado anteriormente, el cual hemos visto que será particularmente bueno cuando l sea notablemente menor que n_p y n_n.

  Véase por ejemplo la Tabla de resultados 1.

  - $\mu_1 = (1, 2)$, $\mu_2 = (10, 11)$, $\Sigma_1 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, $n_1 = n_2 = 300$.

	MST-Class	*MST-RClass_10*	*MST-RClass_50*	*MST-RClass_100*
Min.	0.9444444	0.9722222	0.9777778	0.9722222
1st Qu.	0.9833333	0.9888889	0.9888889	0.9888889
Median	0.9888889	0.9944444	0.9944444	0.9944444
Mean	0.9863889	0.9918889	0.9925000	0.9926111
3rd Qu.	0.9944444	0.9958333	1.0000000	1.0000000
Max.	1.0000000	1.0000000	1.0000000	1.0000000
Elapsed	26.4640000	11.3150000	45.9550000	91.6110000

Tabla 1: Tabla de resultados.

  Se pasa de tener un tiempo de 26,464 segundos con la regla inicial a 91,611 segundos con el método robusto tomando $l = 100$, mientras que con $l = 10$ el tiempo es menor: 11,315 segundos.

- A medida que se aumenta el tamaño de los conjuntos de datos se aprecia más la diferencia (en términos de tiempo) de emplear la regla robusta frente a la original. Aunque tomar solamente 10 submuestras siga proporcionando el menor tiempo, escoger 100 submuestras sigue siendo mucho más rápido que ejecutar el método inicial.

- Mientras que el aumento del tamaño de los conjuntos de datos parece resultar en un aumento exponencial de los tiempos de ejecución para la técnica base, para el método robusto esto no es así, lo cual es mucho más deseable.
- Al pasar de $l = 50$ a $l = 100$ submuestras el tiempo de ejecución se multiplica por 2.

Considerando todas estas observaciones se ha concluido que, aunque el método con $l = 100$ devuelve las mejores clasificaciones, escoger $l = 50$ resulta en tiempos computacionales mucho menores y proporcionando resultados muy próximos. Por tanto, de aquí en adelante se ha escogido $l = 50$, aunque este es un valor para el que hay que tener en cuenta el tamaño del conjunto de datos para evitar situaciones en las que puede que esta cantidad sea demasiado grande o quizás demasiado pequeña.

3.2. Comparaciones con otras reglas

El siguiente paso en el estudio realizado ha sido medir la competitividad de la regla frente a otras reglas como k-vecinos más próximos con $k = 15$ (*knn*), discriminación lineal (*lda*), discriminación cuadrática (*qda*) y la propia regla de Bayes (*bayes*) que, para el caso de dos clases con distribuciones normales, es conocida su expresión explícita. Para comparar estas reglas se han empleado los mismos conjuntos de datos del apartado anterior. Los resultados se han recogido en forma de tablas para su consulta en el Capítulo 1, Sección 1,2 del archivo del repositorio `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`. Se ha añadido además una última columna (*MST-RClass_50*) con los resultados del método robusto con $l = 50$ submuestras (todas ellas con $\sqrt{n}$ observaciones positivas y $\sqrt{n}$ negativas).

Las conclusiones generales obtenidas son las siguientes:

- La regla original obtiene generalmente peores resultados que el resto de métodos, aunque la diferencia no suele ser grande.
- Cuando el resto de métodos se comporta mal *MST-Class* también lo hace, como por ejemplo en configuraciones en las que las medias de ambas clases están muy próximas y tienen la misma matriz de covarianzas. Sin embargo, cuando no se da un solapamiento tan grande se puede observar que el rendimiento de *MST-Class* mejora, lo cual era de esperar.
- La diferencia observada entre la regla *MST-Class* y el resto de clasificadores se ve mitigada empleando *MST-RClass*, siendo incluso en algunos casos mejor. Por ejemplo, veamos los resultados obtenidos con la siguiente configuración (Tabla 2).
  - $\mu_1 = (1, 2)$, $\mu_2 = (2, 7)$, $\Sigma_1 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix}$, $n_1 = n_2 = 300$.

	MST-Class	*knn*	*lda*	*qda*	*bayes*	*MST-RClass_50*
Min.	0.7055556	0.7944444	0.8055556	0.8166667	0.7611111	0.8166667
1st Qu.	0.7666667	0.8555556	0.8611111	0.8611111	0.8319444	0.8611111
Median	0.7916667	0.8722222	0.8722222	0.8833333	0.8555556	0.8777778
Mean	0.7910000	0.8710000	0.8739444	0.8766667	0.8525556	0.8752222
3rd Qu.	0.8166667	0.8902778	0.8888889	0.8902778	0.8777778	0.8944444
Max.	0.8666667	0.9277778	0.9277778	0.9333333	0.9166667	0.9277778

Tabla 2: Tabla de resultados.

Se observa que, en media, pasamos de obtener un porcentaje del 79,1 % de bien clasificados a un 87,52 %, siendo este porcentaje mayor que el del resto de reglas salvo para la discriminación cuadrática.

- A medida que aumentamos el tamaño de las clases puede observarse que, en líneas generales, se reducen muy ligeramente las proporciones para todas las reglas, lo cual puede estar motivado porque estamos clasificando un mayor número de observaciones.

3.3. Pruebas con outliers

Como siguiente paso se ha querido ver el comportamiento de la configuración del método robusto escogida frente a situaciones para las que originalmente fue pensado: conjuntos de datos con datos atípicos o *outliers*. Para ello se han escogido algunas de las configuraciones iniciales para generar los datos, pero a un 4 % del total se les ha cambiado la etiqueta de su clase, es decir, son *outliers* (entendiendo, en este contexto, *outliers* como datos mal etiquetados). En concreto, dichos *outliers* estarán solamente en una de las dos clases. Nuevamente, los resultados se han recogido para su consulta en forma de tablas en el Capítulo 1, Sección 1,3 del archivo del repositorio `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`.

Las conclusiones tras observar los resultados son las siguientes:

- La regla original se ve muy afectada, tal y como se vaticinaba, por los datos atípicos, llegando a clasificar de forma incorrecta en casi la mitad de las ocasiones.
- El método robusto por su parte consigue mejorar estos resultados, proporcionando casi siempre porcentajes de bien clasificados competitivos con otras reglas. Véase por ejemplo el siguiente conjunto de resultados recogidos en la Tabla 3 para la siguiente configuración:
  - $\mu_1 = (1, 2)$, $\mu_2 = (10, 3)$, $\Sigma_1 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, $n_1 = n_2 = 300$.

	MST-Class	*knn*	*lda*	*qda*	*bayes*	*MST-RClass*_50
Min.	0.4239130	0.9347826	0.9510870	0.9565217	0.9239130	0.9565217
1st Qu.	0.4945652	0.9728261	0.9782609	0.9728261	0.9565217	0.9728261
Median	0.5190217	0.9782609	0.9836957	0.9782609	0.9619565	0.9782609
Mean	0.5133696	0.9788043	0.9823370	0.9796739	0.9608152	0.9797283
3rd Qu.	0.5326087	0.9891304	0.9891304	0.9891304	0.9673913	0.9836957
Max.	0.5652174	1.0000000	1.0000000	1.0000000	0.9782609	1.0000000

Tabla 3: Tabla de resultados.

Se observa en la Tabla 3 que, de obtener un 51,33696 % de bien clasificados en media se pasa a obtener un 97,97283 %, lo cual es una diferencia considerable.

3.4. Pruebas en grafos

Líneas de metro

Después de realizar el estudio con datos normales se ha procedido a estudiar otros conjuntos de datos con naturalezas distintas. Dado que *MST-Class* está basado en árboles de expansión mínima, la primera propuesta que se barajó fue considerar datos cuyas clases definieran, de forma natural, grafos. Es por esto por lo que se han considerado en primera instancia las redes de metro de Madrid. En particular, se han considerado diferentes líneas de metro con sus posiciones geográficas como variables explicativas. El objetivo entonces sería clasificar una ubicación dada en alguna de las líneas consideradas. Para ello se han considerado las diferentes paradas de cada línea de metro de Madrid y, para disponer de un conjunto de datos más denso para cada clase, se generaron de forma aleatoria más paradas. Dichas paradas se situaron en las rectas que unen dos paradas consecutivas.

En esta ocasión, dado que no generamos los datos según una distribución dada, las 100 ejecuciones para cada par de líneas de metro se diferencian en la aleatoriedad del conjunto de entrenamiento escogido (que sigue siendo del 70 %). Una vez generados los datos, el estudio realizado fue el siguiente:

- Se comparó la versión robusta de *MST-RClass* con la regla de *knn*, escogiendo para esta última distintos números de vecinos.

  También se comparó con las reglas de discriminación lineal (*lda*) y discriminación cuadrática (*qda*) que asumen normalidad en la distribución de las clases. Cabe mencionar que para la

versión robusta (*MST-RClass*) se escogió $l = 50$ submuestras, todas ellas con $3\sqrt{n}$ observaciones positivas y $3\sqrt{n}$ negativas (recordemos que $n = \text{mín}\{n_p, n_n\}$). Por tanto, cuando nos refiramos a *MST-RClass* sin especificar nada más estaremos haciendo referencia a dicha configuración.

- Asimismo se realizó un estudio comparativo entre diferentes configuraciones del método robusto, variando el número de submuestras y su tamaño, pero manteniendo la complejidad computacional.
- Por último, se hicieron representaciones gráficas de las zonas en las que la configuración robusta *l=50,n=3sqrt* y *knn* con 15 vecinos daban lugar a errores de clasificación.

Todos los resultados se pueden consultar en el Capítulo 2, Sección 2,1, del archivo del repositorio `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`.

Las conclusiones son las siguientes:

- A medida que se aumenta el número de vecinos para *knn* se observa un peor comportamiento, siendo entonces la regla de un vecino más próximo la que proporciona mejores resultados.
- Frente a la opción de *knn* con un vecino obtenemos resultados muy semejantes, siendo mejor *MST-RClass* (en media) en la mitad de las ocasiones.
- La discriminación lineal (*lda*) y la discriminación cuadrática (*qda*) obtienen siempre peores resultados que *MST-RClass* y que *knn* para cualquier número de vecinos, lo cual era de esperar al no estar bajo las hipótesis en las que se sustentan dichos métodos.
- Del análisis comparativo entre las diferentes configuraciones para *MST-RClass* concluimos que la regla es mucho menos sensible que *knn* a cambios en sus parámetros, a pesar de que sí que se observen algunas diferencias.
- Parece tener un mayor impacto el tamaño de las submuestras que el número de las mismas en las proporciones obtenidas. Se observa que a mayor tamaño de las submuestras el porcentaje de bien clasificados aumenta, pero se trata de un porcentaje pequeño, del 1 % como mucho.
- La configuración de *MST-RClass* que devuelve mayores proporciones para cada caso es siempre mejor que *knn* con un único vecino.
- Las zonas en las que se observa que tanto *knn* con $k = 15$ como *MST-RClass* cometen errores de clasificación son las intersecciones de las líneas de metro. Además, parece que los puntos mal clasificados por *knn* dibujan una cruz sobre el plano.

Trayectorias de aviones

En una segunda instancia se han considerado diferentes trayectorias de aviones. Sus posiciones geográficas se han empleado como variables explicativas y como clases los identificadores de los aviones. El objetivo de la clasificación es entonces determinar, dada una coordenada, a qué avión pertenece ese punto de la trayectoria. El repositorio empleado se encuentra en Ghosh et al. (2021). Los datos de las trayectorias fueron recopilados entre el 18 de septiembre de 2020 y el 23 de abril de 2021 en el Aeropuerto Regional de Pittsburgh-Butler, situado al norte de la ciudad de Pittsburgh, Pensilvania. Además, los datos fueron segmentados en *frames*. Un *frame* comienza cuando al menos un avión está activo o entra en el umbral de detección y termina cuando todos los aviones han abandonado las inmediaciones o están inactivos. Para nuestro estudio se han seleccionado *frames* que involucran tan solo a una aeronave.

De nuevo, dado que no generamos los datos según una distribución dada, las 100 ejecuciones para cada par de trayectorias se diferencian en la aleatoriedad del conjunto de entrenamiento escogido (que sigue siendo del 70 %).

El estudio computacional realizado es exactamente el mismo que el efectuado para las líneas de metro. Los resultados se pueden consultar en el Capítulo 2, Sección 2,2 del archivo del repositorio `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`.

Como conclusiones obtenemos que:

- De nuevo, se observa que la mejor opción para *knn* es escoger un único vecino.
- Tanto la discriminación cuadrática como la discriminación lineal obtienen resultados muy por debajo de *MST-RClass* y *knn* para cualquiera de sus configuraciones.
- *MST-RClass* ($l = 50$ con $3\sqrt{n}$ observaciones positivas y $3\sqrt{n}$ negativas) es mejor que *knn* con un único vecino (en media) en un tercio de las ocasiones, siendo los resultados para ambas técnicas más distantes en general que lo observado para las líneas de metro.
- Una vez más, se aprecia que variar los parámetros de *MST-RClass* no conlleva grandes cambios en los resultados como sí ocurre con *knn*. De todas formas, los cambios observados son suficientes para que, nuevamente, exista alguna configuración para *MST-RClass* que obtiene mejores resultados que *knn* con un vecino para todos los pares de trayectorias considerados.
- Las zonas en las que *MST-RClass* y *knn* cometen errores de clasificación son aquellas en las que las trayectorias de ambos aviones están próximas.

3.5. Pruebas en toros

El siguiente paso realizado en el estudio computacional fue considerar toros para generar las clases. Esto fue motivado por la idea de extender lo visto para grafos en la sección anterior. Las características de los toros nos hacían pensar que obtendríamos un buen comportamiento para *MST-RClass* al igual que ocurría con las líneas de metro y las trayectorias de aviones y, al mismo tiempo, tendríamos una distribución subyacente conocida para cada clase. De esta forma, se generaron 100 conjuntos de datos para cada una de las 8 configuraciones diferentes escogidas para los toros. Cada toro generado (que no es lo mismo que cada clase ya que algunas configuraciones consideradas tienen más de un toro en la misma clase) contiene 1000 coordenadas.

El análisis llevado a cabo es el mismo que el efectuado para las líneas de metro y las trayectorias de avión, con la salvedad de que en esta ocasión no se han podido representar los puntos mal clasificados por las reglas, ya que en las 100 ejecuciones para cada configuración los datos varían (lo que permanece igual es la distribución). Los resultados pueden ser consultados en su totalidad en el Capítulo 3, Sección 3,1 del archivo del repositorio `https://nubeusc-my.sharepoint.com/:f:/g/personal/iria_rodriguez_acevedo_rai_usc_es/EhBgIE6rW85JsgwaDfEKDz4BCzOjo1QEUA7GNf-kZVcKJA?e=d8eXs0`.

Las conclusiones obtenidas difieren ligeramente de las de los datos con grafos:

- Para *knn* no se obtiene que la regla con un único vecino devuelva los mejores resultados, sino que dependiendo de la configuración la mejor elección para el número de vecinos varía.
- Los resultados para *MST-RClass* son muy buenos, siendo mejores que *knn* para cualquier k en casi todas las situaciones.

4. CONCLUSIONES

En este trabajo se propone un nuevo algoritmo de clasificación, *MST-Class*, basado en árboles de expansión mínima. Además, se presenta una versión robusta del mismo, denominada *MST-RClass*. La principal conclusión extraída del estudio computacional realizado es que *MST-RClass* obtiene buenos resultados, y se muestra especialmente competitivo con *knn* cuando se aplica a conjuntos de datos cuya naturaleza define grafos en cada clase. Además, *MST-RClass* permite reducir considerablemente el tiempo computacional de las ejecuciones con respecto a *MST-Class*, realizando a su vez mejores clasificaciones, no solo en presencia de *outliers*. Otra de las conclusiones importantes es que la versión robusta es poco sensible a cambios en sus parámetros, al contrario de lo que hemos visto que ocurre con *knn* cuando variamos el número de vecinos.

Como trabajo futuro, el objetivo principal será demostrar la consistencia teórica del método *MST-RClass*, ya que esta propiedad es de especial relevancia para un algoritmo de clasificación. Por otra parte, se consideran otras posibles versiones de *MST-RClass*. Por ejemplo, en esta memoria todas las variables explicativas han tenido el mismo peso en la clasificación, cuando realmente puede darse el caso en el que algunas sean más informativas que otras. Así, podría llevarse a cabo algún tipo de selección de variables antes de aplicar el procedimiento de *MST-RClass*. Aunque

se podrían emplear técnicas estándar de selección de variables, se pretende introducir un nuevo procedimiento basado también en árboles de expansión mínima.

Las conclusiones obtenidas en este trabajo sientan las bases para investigaciones futuras y mejoras en el algoritmo *MST-RClass*, lo que podría ampliar su aplicabilidad a diversos escenarios y problemas de clasificación.

REFERENCIAS

Adler D., Bolker B., Csárdi G., Demont Y, Eddelbuettel D., Fernandez i Marin X., Gebhardt A., Helffrich G., Krylov I., Ming C., Murdoch D., Oleg N., Ooms J., R Core Team, Senger A., Stein M., Strzelecki A., Sumner M., The authors of knitr, The authors of Shiny, Ulrich J., Urbanek S. (2023) *rgl: 3D Visualization Using OpenGL.* R package version 1.1.3. `https://cran.r-project.org/web/packages/rgl/index.html`

Boruvka O. (1926). O jistém problému minimálním (About a certain minimal problem), (3) 37-58 (Czech, German summary).

Boser B. E., Guyon I. M., Vapnik V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chang W., Dunnington D., Henry L., Lin Pedersen T., Takahashi K., Wickham H., Wilke C., Woo K., Yutani H. (2023) *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* R package version 3.4.2. `https://cran.r-project.org/web/packages/ggplot2/index.html`

Cortes C., Vapnik V. (1995). Support vector machine. Machine learning, 20(3), 273-297.

Csárdi G., Horvát S., Müller K., Nepusz T., Noom D., Salmon M., Traag V., Zanini F. (2023). *igraph: Network Analysis and Visualization.* R package version 1.4.2. `https://cran.r-project.org/web/packages/igraph/index.html`

Fisher R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

Fix E., Hodges J. (1951). Discriminatory analysis. Nonparametric discrimination:Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Kruskal J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50.

Ghosh S., Moon B., Oh J., Patrikar J., Scherer S. (2021): TrajAir: A General Aviation Trajectory Dataset. Carnegie Mellon University. Dataset. `https://doi.org/10.1184/R1/14866251.v1`

Prim R. C. (1957). Shortest connection networks and some generalizations. The Bell System Technical Journal, 36(6), 1389-1401.

Quinlan J. R. (1979). Discovering rules by induction from large collections of examples. Expert systems in the micro electronics age.

Quinlan J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

R Core Team (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Rosenblatt F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Zahn C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on computers, 100(1), 68-86.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Estimation of distance correlation: a simulation-based comparative study

Blanca E. Monroy-Castillo[1], Amalia Jácome[1,2] and Ricardo Cao[1,3]

[1]Centro de Investigación CITIC, Universidade da Coruña, A Coruña, 15071, Spain
[2]Faculty of Sciences, Universidade da Coruña, A Coruña, 15071, Spain
[3]Faculty of Computer Sciences, Universidade da Coruña, A Coruña, 15071, Spain

ABSTRACT

Distance correlation is a novel class of multivariate dependence coefficients applicable to random vectors of arbitrary dimensions, not necessarily equal. What sets distance correlation apart is that, unlike Pearson's correlation coefficient, it equals zero if and only if the random vectors are independent. Since its introduction, distance correlation has found numerous applications in various fields, such as variable selection. An estimator of the distance covariance was proposed by Székely *et al.* (2007), which turns out to be an asymptotically unbiased V-statistic. In the same way, an unbiased version of the squared sample distance covariance was proposed and studied in Székely and Rizzo (2014). Finally, Huo and Székely (2016) showed that the unbiased estimator turns out to be a U-statistic. In this study, a simulation is conducted to compare the both distance correlation estimators derived from both distance covariance estimators. The study evaluates efficiency (mean squared error) and compares computational times for both methods under different dependence structures. Also, a new approach, given by a convex combination of the former estimators, is proposed and studied.

Keywords: Distance correlation; U-statistic; V-statistic; simulation study

1. INTRODUCTION

The concept of dependence among random observations plays a central role in many fields, such as statistics, medicine, biology, engineering, and more. Due to the complexity of understanding dependencies in its whole, the strength of dependence is often condensed into a single numerical value known as the correlation coefficient. While there exist several types of correlation coefficients, the most well-known is Pearson's correlation coefficient, which for random variables X and Y with finite variances is defined as follows:

$$\operatorname{cor}(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}},$$

where $\operatorname{Cov}(X,Y)$ represents the covariance of X and Y, and $\operatorname{Var}(Z)$ denotes the variance of any random variable Z. Alternatively, there are other measures of dependence, such as rank correlation, which assesses the relationship between the rankings of two variables or two rankings of the same variable. Examples of rank correlation measures include Spearman's rank correlation coefficient, the Kendall tau correlation coefficient, and Goodman and Kruskal's gamma.

In recent years, a novel measure of dependence for random vectors has been proposed: *distance correlation*. The concept of distance correlation was introduced by Székely *et al.* (2007), who highlight that distance covariance and distance correlation are analogous to product-moment covariance and correlation. However, unlike the classical definition of correlation, distance correlation is zero only when the random vectors are independent. In essence, for all distributions with finite first moments, distance correlation ($\mathcal{R}$) extends the concept of correlation in two fundamental ways:

(i) $\mathcal{R}(X,Y)$ is defined for X and Y in arbitrary dimensions;

(ii) $\mathcal{R}(X,Y)=0$ characterizes independence of X and Y.

Distance correlation satisfies $0 \leq \mathcal{R} \leq 1$, and $\mathcal{R}=0$ if and only if X and Y are independent.

Over the past few years, the concept of distance correlation has undergone significant study and extension. For instance, Székely and Rizzo (2009) introduced the idea of covariance in relation to Brownian motion, leading to the concept of Brownian distance covariance. Meanwhile, the uniqueness of distance covariance is demonstrated by Székely and Rizzo (2012). Additionally, Székely and Rizzo (2013) explored the application of distance correlation to the problem of testing independence of high-dimensional random vectors. Likewise, the notion of partial distance correlation was introduced by Székely and Rizzo (2014). So, Székely *et al.* (2007) proposed a sample distance covariance estimator, and they proved that the estimator is a V-statistic. In Székely and Rizzo (2014), intermediate results are presented, which yield an unbiased estimator of the squared distance covariance. The unbiased estimator is established as a U-statistic in Huo and Székely (2016), along with the development of a new algorithm to compute it. This algorithm boasts a computational complexity of $\mathcal{O}(n \log n)$, a significant improvement over the $\mathcal{O}\left(n^2\right)$ complexity incurred by direct implementation of the V-estimator proposed by Székely *et al.* (2007).

It is worth noting that, as far as we are aware, the advantages and disadvantages associated with each estimator of the distance correlation (U-estimator and V-estimator) have not been extensively explored in the existing literature. For instance, Edelmann *et al.* (2022) utilized the U-estimator to propose an extension suitable for right-censored data, while Wang *et al.* (2015) introduced an extension for conditional distance correlation, employing both estimators in their approach.

In this work, both estimators are compared, and a new approach is proposed. The paper is organized as follows. Section 2 presents the estimators together with their main properties. A simulation study is conducted in Section 3 in which the effectiveness of the methods is evaluated by comparing their mean squared error (MSE), bias and variance. In addition, a method is proposed to solve the problem of negative values that might arises when using the U-estimator. In addition, a new approach consisting of a convex combination of these estimators is presented and its effectiveness is evaluated.

2. PRELIMINARIES

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors, where p and q are positive integers. The characteristic function of X and Y are denoted as ϕ_X and ϕ_Y, respectively, and the joint characteristic function of X and Y is denoted as $\phi_{X,Y}$. So, the squared distance covariance between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{V}^2(X,Y)$ defined by

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|\phi_{X,Y}(t,s)-\phi_X(t)\phi_Y(s)\|^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(t,s)-\phi_X(t)\phi_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}}dt\,ds,\end{aligned}$$

where $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$. Similarly, distance variance is defined as the square root of

$$\mathcal{V}^2(\mathrm{X}) = \mathcal{V}^2(X,X) = \|\phi_{X,X}(t,s)-\phi_X(t)\phi_X(s)\|^2.$$

Finally the distance correlation ($\mathcal{R}$) between random vectors X and Y with finite first moments is the positive square root of the nonnegative number $\mathcal{R}^2(X,Y)$ defined by

$$\mathcal{R}^2(X,Y) = \begin{cases}\frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y)>0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y)=0.\end{cases}$$

On the other hand, an equivalent form to compute the distance covariance through expectations is proposed in Remark 3 of Székely *et al.* (2007). This is, if $E|X|_p^2 < \infty$ and $E|Y|_q^2 < \infty$, then $E[|X|_p|Y|_q] < \infty$, and

$$\begin{aligned}\mathcal{V}^2(X,Y) &= E[|X_1-X_2|_p|Y_1-Y_2|_q]+E[|X_1-X_2|_p]E[|Y_1-Y_2|_q] \\ &\quad -2E[|X_1-X_2|_p|Y_1-Y_3|_q], \qquad (1)\end{aligned}$$

where $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) are independent and identically distributed as (X, Y).

For an observed random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ from the joint distribution of random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, Székely *et al.* (2007) proposed the empirical distance covariance ($\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$) as follows. The empirical distance covariance $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ is the positive square root of the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl}, \tag{2}$$

where A_{kl} and B_{kl} denote the corresponding double-centered distance matrices defined as

$$A_{kl} = \begin{cases} a_{kl} - \frac{1}{n}\sum_{j=1}^{n} a_{kj} - \frac{1}{n}\sum_{i=1}^{n} a_{il} + \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}, & k \neq l \\ 0, & k = l, \end{cases}$$

where $a_{kl} = |X_k - X_l|$ are the pairwise distances of the X observations. The terms B_{kl} are defined similarly but using $b_{kl} = |Y_k - Y_l|$. In the same way

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2. \tag{3}$$

Theorem 1 in Székely *et al.* (2007) proves that $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \geq 0$. Moreover, it is proved that under independence, $\mathcal{V}_n^2$ is a degenerate kernel V-statistic. The estimator has a computational complexity of $\mathcal{O}(n^2)$. Therefore, the empirical distance correlation $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ is the square root of

$$\mathrm{dCorV}^2(\mathbf{X}, \mathbf{Y}) = \mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}}, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0 \\ 0, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) = 0, \end{cases} \tag{4}$$

which is always non-negative.

In the same way, in Székely and Rizzo (2014) it is defined the $\mathcal{U}-$ centered matrix as follows. Let $A = (a_{ij})$ be a symmetric, real valued $n \times n$ matrix with zero diagonal, $n > 2$. The (i, j)th entry of the $\mathcal{U}-$centered matrix $\tilde{A}$ is defined by

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2}\sum_{l=1}^{n} a_{il} - \frac{1}{n-2}\sum_{k=1}^{n} a_{kj} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{kl}, & i \neq j; \\ 0, & i = j. \end{cases}$$

Here "$\mathcal{U}-$centered" is so named because the inner product,

$$\mathcal{U}_n^2(\mathbf{X}, \mathbf{Y}) = \left(\tilde{A} \cdot \tilde{B}\right) = \frac{1}{n-3} \sum_{i \neq j} \tilde{A}_{ij} \tilde{B}_{ij}, \tag{5}$$

defines an unbiased estimator of the squared distance covariance. Huo and Székely (2016) proved that the estimator in Equation (5) is a U-statistic. This reformulation allowed the development of a fast algorithm that can be implemented with a computational complexity of $\mathcal{O}(n \log n)$.
Thus, it is possible to define the empirical distance correlation through U-statistics (dCorU) which is the square root of

$$\mathrm{dCorU}^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{U}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y})}}, & \mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y}) > 0 \\ 0, & \mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y}) = 0, \end{cases} \tag{6}$$

where $\mathcal{U}_n^2(\mathbf{X})$ represents the distance variance of $\mathbf{X}$, similarly $\mathcal{U}_n^2(\mathbf{Y})$ for $\mathbf{Y}$.

Based on the previous findings, several software packages have been created. They are available for use in the R software environment (R Core Team, 2022) and Python (Van Rossum and Drake Jr, 1995). A comparative analysis of the different packages across both programming languages is

conducted in Ramos-Carreño and Torrecilla (2023). Additionally, an open source Python package dedicated to distance correlation and other statistics is introduced, the **dcor** package (Ramos-Carreño, 2022). In the Python domain, the examined libraries consist of **statsmodels** (Seabold and Perktold, 2010), **hyppo** (Panda *et al.*, 2021), and **pingouin** (Vallat, 2018). In the R environment, the compared packages encompass **energy** (Rizzo and Székely, 2022) and **dcortools** (Edelmann and Fiedler, 2022). Also, the **Rfast** package (Papadakis *et al.*, 2022) provides functions such as dvar, dcov, dcor, and bcdcor. Notably, the rapid method introduced by Huo and Székely (2016) is employed for the computation of vector distance variance. Meanwhile, the bcdcor function facilitates the calculation of bias-corrected distance correlation for two matrices.

3. SIMULATION STUDY

A Monte Carlo simulation study was conducted in order to compare the efficiency of the dCorU and dCorV estimators under different dependence structures. For the simulation study, the **dcortools** package is employed, specifically utilizing the distcor function. To calculate the distance correlation using U-statistics (dCorU), the code used is distcor(X, Y, bias.corr = TRUE). Conversely, to compute dCorV, the code is distcor(X, Y, bias.corr = FALSE) or simply distcor(X, Y). To facilitate a comprehensive comparison efficiency (MSE) and computational time of the estimators for three distinct models are utilized, Farlie-Gumbel-Morgenstern model (FGM), a bivariate normal model and a nonlinear model. These models encompass varying levels of dependence and are evaluated across three sample sizes, 100, 1000, and 10000. Each simulation is performed with 1000 Monte Carlo repetitions.

3.1 FGM-Model

One of the most popular parametric families of copulas is the Farlie-Gumbel-Morgenstern (FGM) family that is defined by

$$C^{FGM}(u,v) = uv[1+\theta(1-u)(1-v)], \qquad \theta \in [-1,1]$$

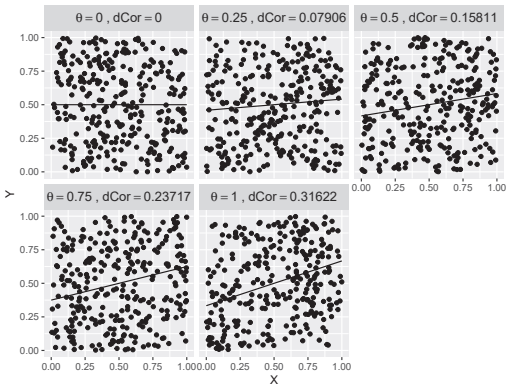
with copula density given by

$$c^{FGM}(u,v) = 1+\theta(2u-1)(2v-1), \qquad \theta \in [-1,1]. \tag{7}$$

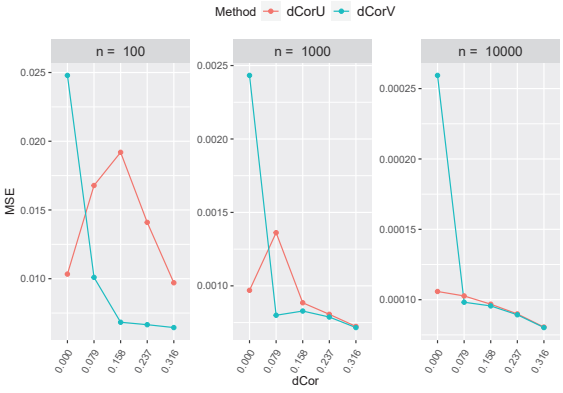
A well-known limitation of this family is that it does not allow the modeling of high dependences since Pearson's linear correlation coefficient is limited to $\rho = \frac{\theta}{3} \in [-0.33, 0.33]$.

The precise computation of distance covariance is achieved using Equation (1), wherein the density function corresponds to Equation (7). Similarly, the calculations for $\mathcal{V}(X)$ and $\mathcal{V}(Y)$ were performed. As a consequence, the $\mathcal{R} = \frac{\theta}{\sqrt{10}}$ is obtained.

Five samples drawn from different values of θ are shown in Figure 1a, together with the value of θ and the corresponding distance correlation. The lines represent the conditional mean $E[Y|X=x]$. It is evident that the dependence is relatively weak, falling within the range of $0 \leq \mathcal{R} \leq 0.31622$. For each scenario, metrics such as mean, variance, mean squared error (MSE), and bias were computed. The comparison between the MSE values obtained with dCorU and dCorV is presented in Figure 1b. When the sample size is $n = 100$, the difference between the two estimators is very clear across all θ values. In cases of dependence, dCorV outperforms dCorU in terms of MSE for all three scenarios. When the sample size is small, dCorV offers a more accurate estimation, except in the scenario of complete independence. As the sample size increases, both estimators perform similarly, except in the case of independence. This implies that for the FGM-Model under independence, dCorU beats dCorV.



(a) FGM samples ($n = 300$) for different values of θ and their corresponding distance correlation.



(b) MSE of dCorU and dCorV under the FGM-Model for $n = 100, 1000, 10000$ for different distance correlation values, from $\mathcal{R} = 0$ ($\theta = 0$) to $\mathcal{R} = 0.316$ ($\theta = 1$), computed with 1000 samples.

Figure 1: FGM-Model.

In Figure 2, the bias and variance are displayed for each sample size ($n = 100, 1000, 10000$). It is apparent that both estimators exhibit significant similarity when the sample size is larger ($n \geq 1000$). Conversely, in cases of smaller sample sizes, dCorV boasts lower variance, while dCorU displays a negative bias that converges to zero as dependence strengthens.

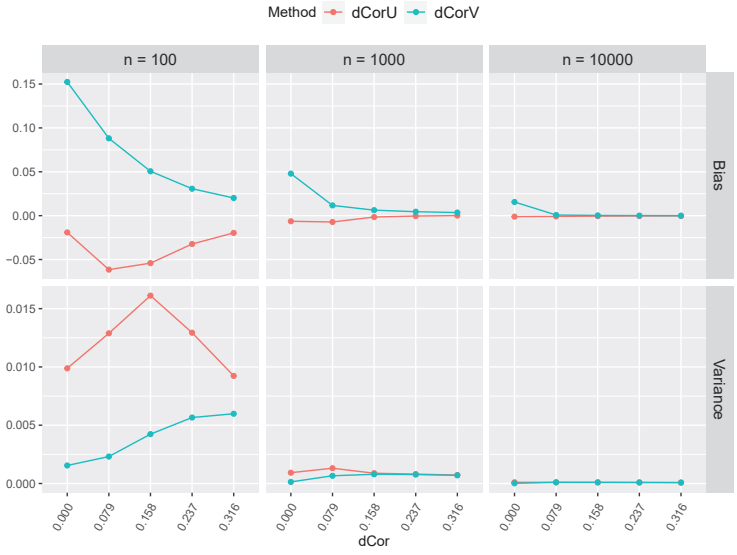


Figure 2: Bias and variance under the FGM-Model with the different sample sizes and distance correlation.

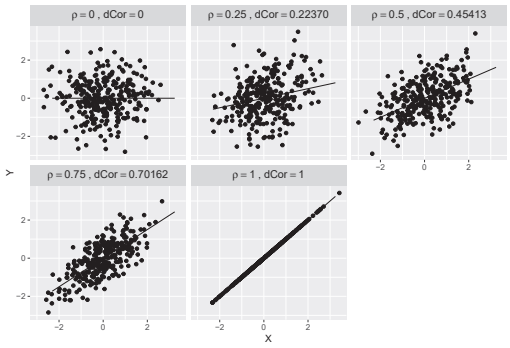
3.2 Bivariate normal model

If (X, Y) has bivariate normal distribution with unit variance each and with Pearson's correlation coefficient $\rho = \rho(X, Y)$, Székely *et al.* (2007) prove that the distance correlation is given by the square root of

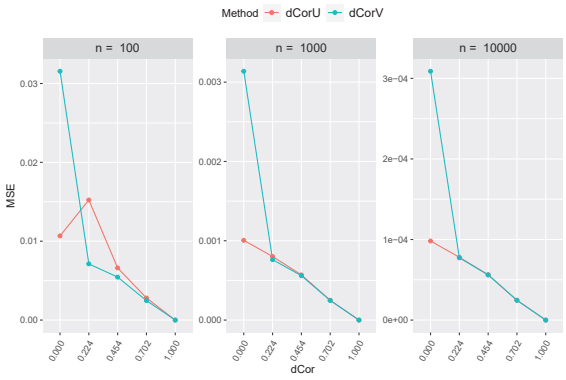
$$\mathcal{R}^2(X, Y) = \frac{\rho \arcsin \rho + \sqrt{1 - \rho^2} - \rho \arcsin \rho/2 - \sqrt{4 - \rho^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

For the bivariate normal (BN) model five samples for different values of ρ are shown in Figure 3a.

This model accommodates complete dependence, indicated by $\mathcal{R} = 1$ when $\rho = 1$. This configuration provides an opportunity to observe the behavior of the estimators in an scenario of strong dependence. Under conditions of independence, the optimal estimation is achieved using dCorU.



(a) Bivariate normal model samples ($n = 300$) for different values of ρ and their corresponding distance correlation.



(b) MSE of dCorU and dCorV under the BN-model for different distance correlation values, from $\mathcal{R} = 0$ ($\rho = 0$) to $\mathcal{R} = 1$ ($\rho = 1$), computed with 1000 samples.

Figure 3: Bivariate normal model.

When dependence becomes more pronounced, specifically for $\mathcal{R} \geq 0.45$, both estimators exhibit similar performance across all three sample sizes (100, 1000, and 10000). In fact, when the sample size is either 1000 or 10000, and $\mathcal{R} \geq 0.22$, the two estimators are virtually indistinguishable in terms of MSE.

A similar pattern is also evident in Figure 4, which depicts how both the bias and variance tend to align for instances of substantial dependence ($\mathcal{R} \geq 0.454$). For $n = 100$, dCorV beats dCorU except for independence or extreme independence. Furthermore, for larger sample sizes ($n \geq 1000$), both estimators yield very close outcomes.

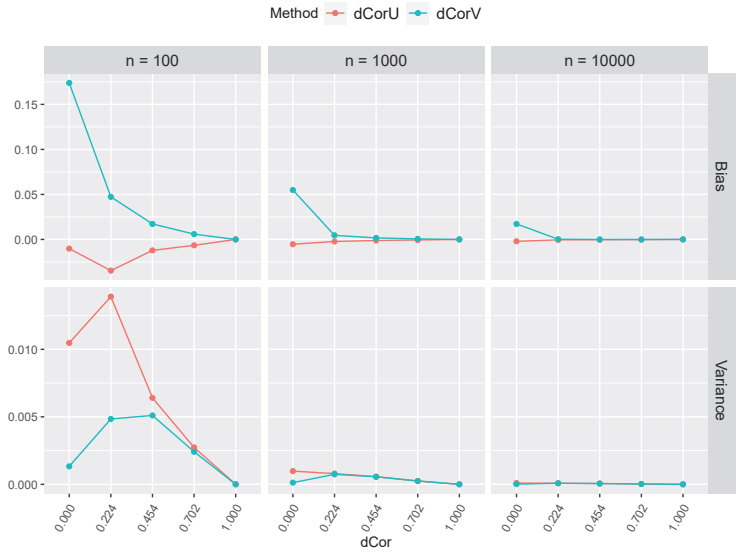
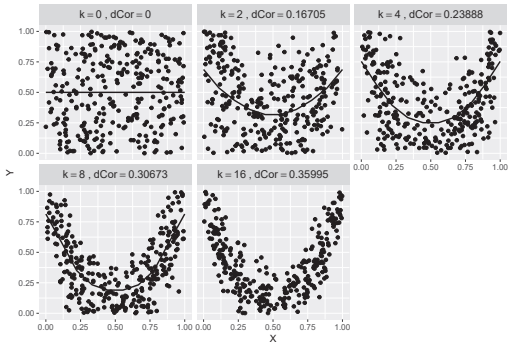


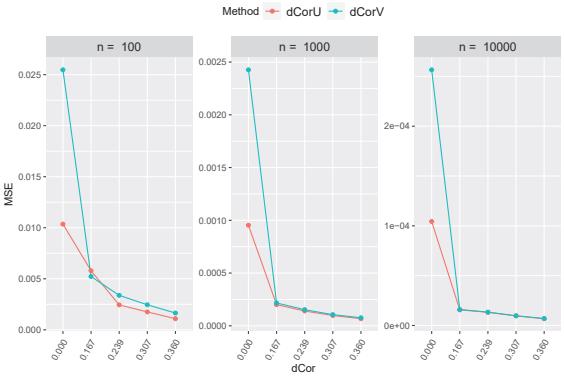
Figure 4: Bias and variance under the bivariate normal model for different sample sizes and distance correlation.

3.3 Nonlinear model

Consider (X, Y) sampled from the joint density function $f_{X,Y}(x, y)$, defined as follows: $f_{X,Y}(x, y) = c\left[1 - \left(y - 4\left(x - \frac{1}{2}\right)^2\right)^2\right]^k I_{[0,1]}(x) I_{[0,1]}(y)$, where $k \in \mathbb{N}$ and c is a constant that depends on the value of k. In this model, the values of k employed are 0, 2, 4, 8, and 16. The computation of the distance correlation utilizes the expression provided in Equation (1).The resulting distance correlations for each sample are illustrated in Figure 5a.



(a) Nonlinear model samples ($n = 300$) for different values of k and their corresponding distance correlation.



(b) MSE of dCorU and dCorV under the nonlinear model for $n = 100, 1000, 10000$ for different distance correlation values, from $\mathcal{R} = 0$ ($k = 0$) to $\mathcal{R} = 0.36$ ($k = 16$), computed with 1000 samples.

Figure 5: Nonlinear model.

It is important to note that, in this model, the maximum value of the distance correlation is not 1 but $\mathcal{R} = 0.41$, achieved when X and Y are totally dependent (i.e. when $k \to \infty$). Furthermore, with $k = 64$, the value of $\mathcal{R}$ is 0.406. So, unlike the linear models (FGM and normal bivariate), low levels of $\mathcal{R}$ ($0.22 \leq \mathcal{R} \leq 0.41$) could indicate a strong relationship between the X and Y variables. When considering independence, the superior performance is obtained using dCorU.

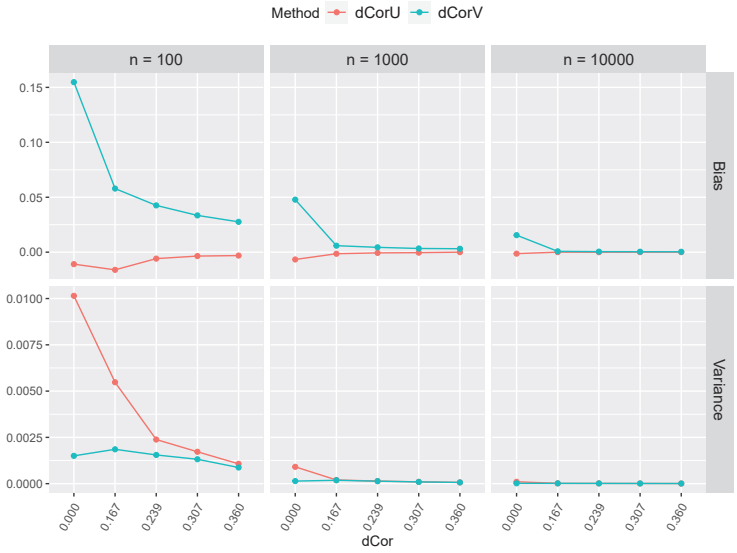


Figure 6: Bias and variance under the nonlinear model with different sample sizes and values of $\mathcal{R}$.

The most substantial disparity with the linear models becomes apparent when the sample size is 100, wherein dCorU offers a more accurate estimation. These distinctions are also visually represented in Figure 6, which illustrates the bias and variance outcomes. Finally, the computational time for each estimator are in Table 1. The characteristics of the computer equipment used are the following ones: CPU 12th Gen Intel(R) Core(TM) i7-1280P 2.00 GHz and RAM 16 GB.

	$n = 100$		$n = 1000$		$n = 10000$	
	dCorU	dCorV	dCorU	dCorV	dCorU	dCorV
Time	0.36	0.33	0.63	0.58	2.22	2.51

Table 1: Computational time in secs for 1000 samples.

Note that there are no significant differences in the computation times among each of the estimators. It is important to highlight that these computation times were obtained with the **dcortools** package. If the **energy** package were used, the computation times of both methods are not so close.

3.4 New proposal

A remarkable feature found in the simulation study was the occurrence of negative values when computing dCorU^2 in Equation (6) under conditions of independence or weak dependence, especially with small sample sizes. This issue arises due to the fact that $\mathcal{U}_n^2$ is an unbiased estimator of the distance covariance, which is 0 under independence. To address this concern, two alternative approaches are investigated in this study:

- Replace the negative values with their absolute values (dCorU(A)).
- Truncate negative values to zero (dCorU(T)).

Table 2 shows the percentage of negative values obtained when the dCorU^2 estimator is used.

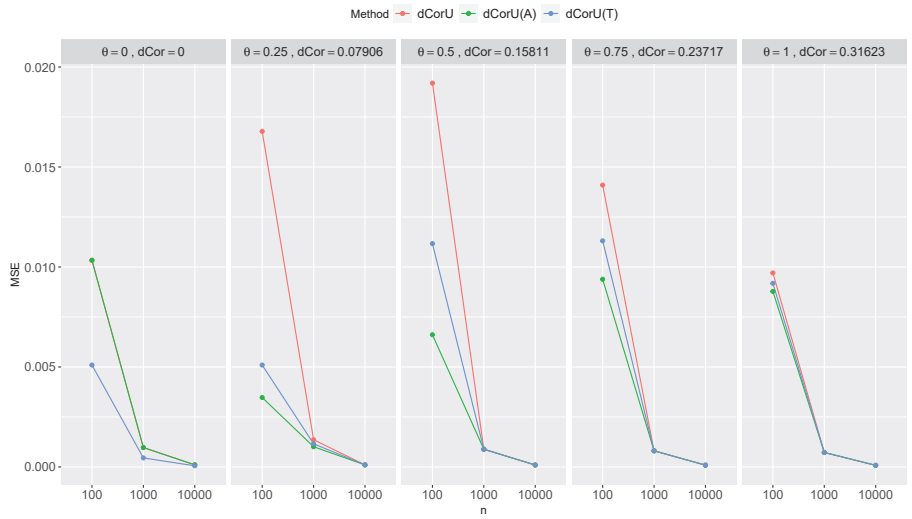
FGM-Model					Bivariate Normal				
		Percentage of negative values					Percentage of negative values		
θ	$\mathcal{R}$	$n = 100$	$n = 1000$	$n = 10000$	ρ	$\mathcal{R}$	$n = 100$	$n = 1000$	$n = 10000$
0	0	65.4	65.6	62.3	0	0	59.2	62.6	64.3
0.25	0.079	50.3	5.1	0	0.25	0.224	9.3	0	0
0.5	0.158	26	0	0	0.5	0.454	0	0	0
0.75	0.237	6.5	0	0	0.75	0.702	0	0	0
1	0.316	1	0	0	1	1	0	0	0

Nonlinear model				
		Percentage of negative values		
k	$\mathcal{R}$	$n = 100$	$n = 1000$	$n = 10000$
0	0	60.6	65.4	62.7
2	0.167	4.7	0	0
4	0.239	0.1	0	0
8	0.307	0	0	0
16	0.360	0	0	0

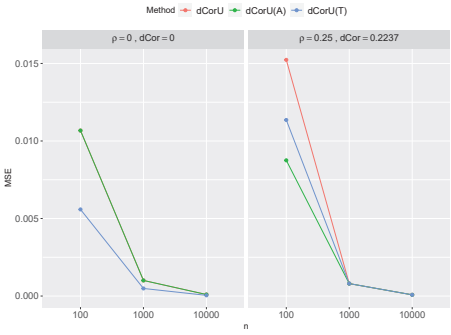
Table 2: Percentage of negative values obtained when computing dCorU^2 for each of the models across different scenarios.

Figure 7 illustrates the MSE of these modified versions of dCorU obtained for each model. Each plot shows the MSE across varying levels of dependence, denoted by dCor, for three distinct sample sizes ($n = 100, 1000, 10000$). When the computation of dCorU^2 yields a negative value, the estimator dCorU is obtained anyway as follows: the sign of dCorU^2 will be considered, and the

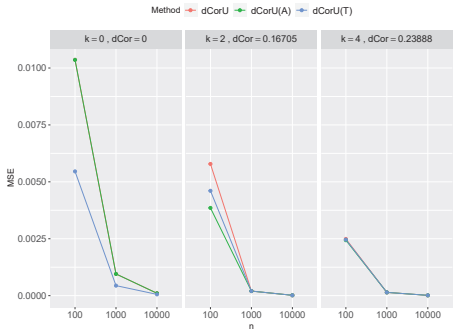
square root of the absolute value will be calculated, i.e. $\text{dCorU} = \text{sign}(\text{dCorU}^2)\sqrt{|\text{dCorU}^2|}$. A few important observations can be made.



(a) FGM-Model.



(b) Bivariate normal model.



(c) Nonlinear model.

Figure 7: MSE obtained for each model with dCorU, dCorU using absolute value with negative (dCorU(A)) and dCorU truncating to zero the negative values (dCorU(T)).

Firstly, within the FGM-Model, dCorU^2 yields some negative results under all the dependence scenarios when $n = 100$ (Figure 7a, Table 2). On the other hand, with the bivariate normal and nonlinear models, dCorU^2 generates negative numbers only under certain scenarios (Figures 7b, 7c, Table 2).

Under independence, the best estimation is obtained using dCorU(T), which involves truncating negative values of dCorU^2 to zero. Obviously, the results with dCorU and dCorU(A) are identical. This is because the squared of the difference between the estimated value and the real value $((\text{dCorU} - \mathcal{R})^2)$ does not deviate from the difference obtained when using the absolute value $((\text{dCorU(A)} - \mathcal{R})^2)$ when $\mathcal{R} = 0$. Under dependence, the best estimator, as indicated by MSE, is dCorU(A), particularly for moderate levels of dependence (i.e. $\mathcal{R} < 0.23$). However, the MSE of the three methods are very similar for the nonlinear model (Figure 7c).

In scenarios of high dependence, such as when $\mathcal{R} > 0.25$, the three estimators have almost the same MSE. The results up to this point are inconclusive, making it challenging to definitively determine the preferred estimator across various scenarios. On one hand, dCorV boasts lower variance and consistently produces positive estimates. On the other hand, dCorU yields superior estimates under independence and displays lower bias. The most pronounced discrepancies are observed in cases of low dependence and small sample sizes. For $n = 10000$, the presence of negative values is

limited to the case where $\mathcal{R} = 0$.

To tackle the estimator selection challenge, a potential solution is proposed: a convex linear combination of one of the dCorU estimators (i.e., dCorU, dCorU(A), or dCorU(T)) and the dCorV estimator:

$$\begin{aligned} \text{LdCorU} &= \lambda\text{dCorU} + (1-\lambda)\text{dCorV}, \\ \text{LdCorU(A)} &= \lambda\text{dCorU(A)} + (1-\lambda)\text{dCorV}, \\ \text{LdCorU(T)} &= \lambda\text{dCorU(T)} + (1-\lambda)\text{dCorV}, \end{aligned}$$

where $\lambda \in [0, 1]$. The optimal value of λ with respect to the MSE for the LdCorU estimator is:

$$\lambda = \frac{-\text{Cov}\left(\text{dCorU},\, \text{dCorV}\right) + \text{Var}\left(\text{dCorV}\right) + \text{Bias}\left(\text{dCorV}\right)\left(\text{Bias}\left(\text{dCorV}\right) - \text{Bias}\left(\text{dCorU}\right)\right)}{\text{Var}\left(\text{dCorU}\right) + \text{Var}\left(\text{dCorV}\right) - 2\text{Cov}\left(\text{dCorU}, \text{dCorV}\right) + \left(\text{Bias}\left(\text{dCorU}\right) - \text{Bias}\left(\text{dCorV}\right)\right)^2}. \tag{8}$$

The optimal value of λ for the LdCorU(A) and LdCorU(T) estimators can be derived similarly. However, in practice, obtaining the exact value of λ is not feasible. To facilitate a comparison in this study, the λ values obtained using Equation (8) are studied alongside an estimator of it using the bootstrap method. The comparison was conducted for a sample size of $n = 100$ across various levels of dependence.

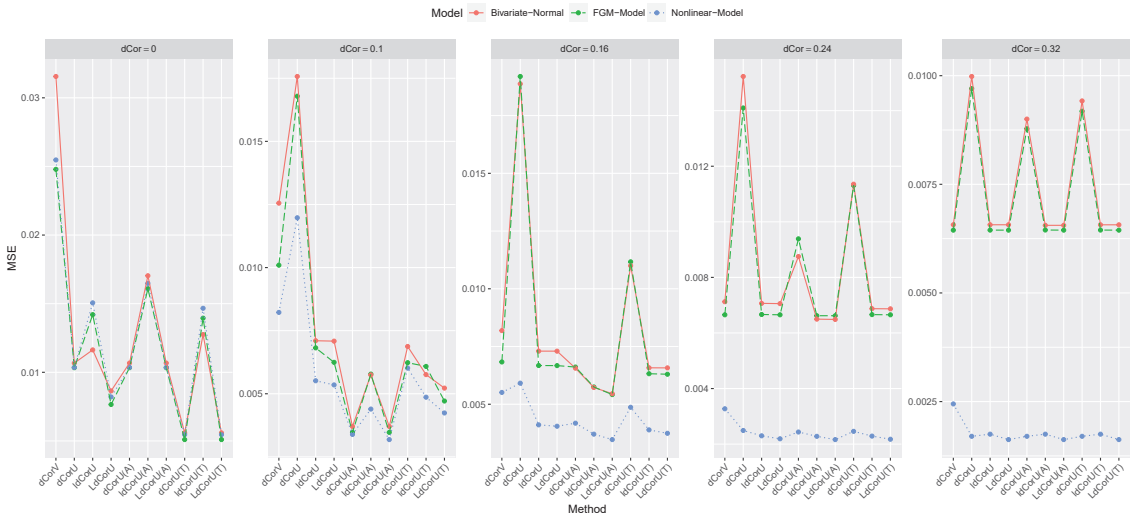
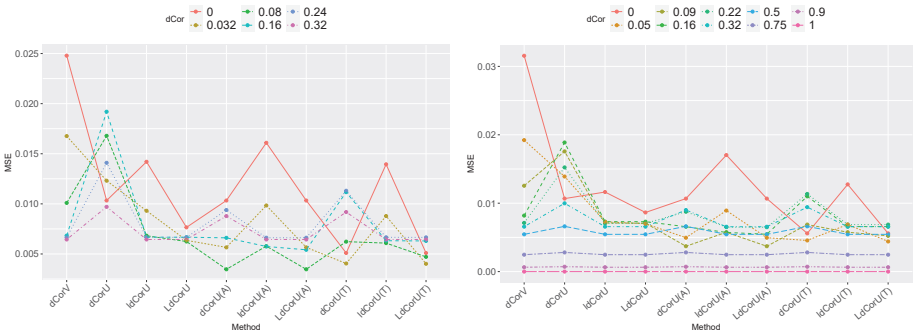


Figure 8: Comparison of the MSE between the different estimators and their respective convex linear combination with $n = 100$. Specifically, ldCorU denotes the combination λdCorU + $(1 - \lambda)$dCorV with λ estimated using 1000 bootstrap replications. Similarly, LdCorU denotes the combination λdCorU + $(1-\lambda)$dCorV using the real optimal value of λ (Eq. (8)). The same naming convention applies to dCorU(A) and dCorU(T), each representing their respective combinations. The values of $\mathcal{R}$ shown are rounded values of the corresponding ones from each model.

The results are shown in the Figure 8. The estimators computed with λ approximated by bootstrap are referred as ldCorU, ldCorU(A) and ldCorU(T) respectively. Indeed, specific scenarios reveal that the lowest MSE is achieved through dCorU(A) or dCorU(T).

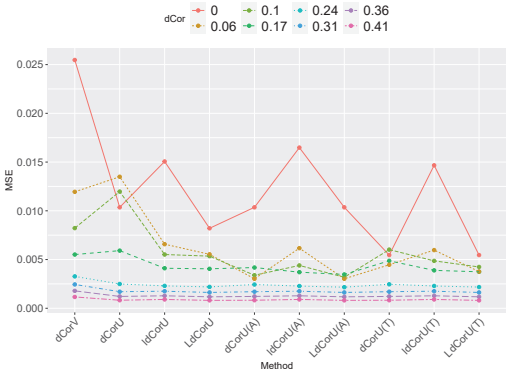
However, a remarkable fact is that the convex combination, employing the value of λ estimated through bootstrapping, consistently provides superior estimates compared to using dCorU and dCorV individually. Moreover, it is worth highlighting that the outcomes for the nonlinear model deviate from those of the other models: the MSE notably decreases as dependence increases.

To show more precise information for each of the estimates across various scenarios, Figure 9 displays the MSE of each model as a function of $\mathcal{R}$ with $n = 100$.



(a) FGM-Model

(b) Bivariate normal model.



(c) Non-linear model.

Figure 9: Comparison of the MSE among the different estimators and their respective convex linear combinations performed for the provided models across various levels of dependence, from independence ($\mathcal{R} = 0$) to the strongest dependence supported by each model for $n = 100$. The λ value was estimated using 1000 bootstrap replications.

Consistent results are observed in all scenarios except the case of independence. In that case, the bootstrap estimates of the lambdas differ from the actual results. It should be noted that under moderate and strong dependency ($\mathcal{R} \geq 0.5$ in the bivariate normal model, $\mathcal{R} \geq 0.17$ for the nonlinear model), there are no notable differences among the estimators or their corresponding combinations.

As previously indicated, Figure 9 also illustrates that in situations involving any form of dependence, even if it is minimal, the convex combination with the estimated lambda, λdCorU + $(1-\lambda)$dCorV, yields a more accurate estimate compared to the original estimators (dCorU, dCorV).

4. CONCLUSIONS

This research addressed the problem of studying the performance of the estimators of the distance correlation dCorU and dCorV by means of Monte Carlo simulation. The results show that the conclusions depend on the scenario. Under independence, dCorU looks preferable over dCorV for all the scenarios analyzed. When considering the proposed versions of the estimator dCorU based on truncating or computing the absolute value, the superior estimation is in general obtained using dCorU(T).

However, under dependence, the conclusions are different. The dCorV estimator aligns with the best results according to the Mean Squared Error (MSE) for the linear models (FGM and Bivariate normal), and also for the nonlinear model under weak dependence. But when considering the given approaches, specifically dCorU(A) and dCorU(T), the optimal estimator seems to be dCorU(A).

In practice, it is difficult to determine whether it is a linear or nonlinear model. Moreover, discerning between total independence and weak dependence can be equally difficult. In such cases, a prudent approach is to employ a convex linear combination, represented as λdCorU + $(1-\lambda)$dCorV, which tends to produce a superior estimate compared to using only dCorU or dCorV individually when the optimal value of λ is estimated via bootstrapping.

Finally, in terms of computational time, both estimators, dCorU and dCorV, are similar. When the convex linear combination is used, the time increases depending on the number of bootstrap replicates used.

REFERENCES

Edelmann, D. and Fiedler, J. (2022). *dcortools: Providing Fast and Flexible Functions for Distance Correlation Analysis*. URL: `https://CRAN.R-project.org/package=dcortools`, R package version 0.1.6.

Edelmann, D., Welchowski, T. and Benner, A. (2022). A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics*, 78(3), 867–879.

Huo, X. and Székely, G. J. (2016). Fast Computing for Distance Covariance. *Technometrics*, 58(4), 435–447.

Panda, S., Palaniappan, S., Xiong, J., Bridgeford, E., Mehta, R. and Shen, C. (2021). *hyppo: A multivariate hypothesis testing Python package*. URL: `https://github.com/neurodata/hyppo.`

Papadakis, M., Tsagris, M., Dimitriadis, M., Fafalios, S., Tsamardinos, I., Fasiolo, M., Borboudakis, G., Burkardt, J., Zou, C., Lakiotaki, K. and Chatzipantsiou., C. (2022). *Rfast: A Collection of Efficient and Extremely Fast R Functions*. URL: `https://CRAN.R-project.org/package=Rfast`, R package version 2.0.6.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Ramos-Carreño, C. (2022). *dcor: distance correlation and energy statistics in Python*. URL: `https://github.com/vnmabus/dcor`.

Ramos-Carreño, C. and Torrecilla, J. L. (2023). dcor: Distance correlation and energy statistics in Python. *SoftwareX*, 22, 101326.

Rizzo, M. L. and Székely, G. J. (2022). *energy: E-Statistics: Multivariate Inference via the Energy of Data*. URL: `https://CRAN.R-project.org/package=energy`, R package version 1.7-11.

Seabold, S. and Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with Python*. URL: `https://github.com/statsmodels/statsmodels/`.

Székely, G. J. and Rizzo, M. L. (2009). Brownian Distance Covariance. *Annals of Applied Statistics*, 3(4), 1236–1265.

Székely, G. J. and Rizzo, M. L. (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters*, 82(12), 2278–2282.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117, 193–213.

Székely, G. J. and Rizzo, M. L. (2014). Partial Distance Correlation with Methods for Dissimilarities. *Annals of Statistics*, 42(6), 2382–2412.

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

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 1026.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

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.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

AN ALLOCATION RULE FOR GRAPH MACHINE SCHEDULING PROBLEMS

Laura Davila Pena[1], Peter Borm[2], Ignacio García-Jurado[3] and Jop Schouten[4]

[1]Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain.
[2]Department of Econometrics & Operations Research, Tilburg University, the Netherlands.
[3]CITMAga, Department of Mathematics, Universidade da Coruña, Spain.
[4]Amsterdam School of Economics, University of Amsterdam, the Netherlands.

ABSTRACT

This work studies a new type of interactive Operations Research problem, called a graph machine scheduling problem (GMS-problem). A GMS-problem combines aspects from minimum cost spanning tree problems and sequencing problems. Given a graph, we aim to first establish a connection order on the players such that the total cost of connecting them to a source is minimal and second to find a cost allocation of such an optimal order among the players involved. We restrict attention to GMS-problems on trees and propose a recursive method to solve these tree GMS-problems integrated with an allocation approach. This latter mechanism consistently and recursively uses myopic reference orders to determine potential cost savings, which will then be appropriately allocated. Interestingly, the transition process from a myopic reference order to an optimal one will be smooth using the switching of blocks of agents based on the basic notion of merge segments.

Keywords: Scheduling; Connection problems; Sequencing problems; Graph machine scheduling problems; Cost allocation.

1. INTRODUCTION

As argued in Bergantiños et al. (2014), there are many real-life problems that require the construction of infrastructures to connect a set of agents to a source, either directly or indirectly. One of them is the urban supply of water from a general reservoir to certain points of interest (agents), which involves building pipelines throughout a city. Installing pipes between two points takes a certain amount of time. The first problem that arises in this kind of situation is the question of where the pipelines should be. The objective will thus be to connect all the agents to the network in such a way that the total time involved is minimized. The construction time can be interpreted as a cost to be minimized. To address this problem, the standard minimum cost spanning tree (MCST) setting has been widely applied (see, for example, Curiel, 1997 or Bergantiños & Lorenzo, 2004). In this way, the focus is on determining so-called MCSTs. A tree is a set of edges such that there is a single path from the source to each of the agents, and the cost of a tree is the sum of the costs of all the edges belonging to it. Other real-world situations that can be modeled as an MCST problem can be found in Claus & Kleitman (1973). MCSTs can be computed in a polynomial time and the most common methods are Kruskal's algorithm and Prim's algorithm. A further issue of relevance in an interactive optimization setting is the allocation of costs among the different agents involved. An adequate allocation will serve to establish and maintain cooperation between the agents. Out of the extensive literature on MCST problems that approach the cost allocation issue, we refer the reader to the recent review of Bergantiños and Vidal-Puga (2021).

Looking back at the urban supply of water, it is often essential for the agents to be provided with water at all times (e.g., a hospital), so they have to contract an external service company for as long as the water supply does not reach them. Thus, each agent has an associated coefficient

that indicates the cost per unit of time in the system, i.e., per unit of time for which the pipes that connect it to the source are not yet constructed. The total construction time will depend on *when* this agent is connected to the source: for example, if agent 2 is connected to the source via agent 1, then the pipeline connecting the source to agent 1 must be constructed first, followed by the pipeline connecting agent 1 to agent 2. Hence, the total time required to connect agent 2 to the source would be the sum of the construction times of the pipelines connecting the source to agent 1, and agent 1 to agent 2. Thus, the objective is to minimize the total aggregate costs instead of just the total construction time for the project as a whole.

Situations such as those described above result in a new type of problem, which we call the graph machine scheduling problem (GMS-problem). One issue we would like to highlight is the proximity of our problem to a sequencing problem, of which we will give a brief description below.

In deterministic one-machine sequencing problems, a set of jobs needs to be processed on a machine. Each of these jobs is identified with one agent and has associated with it: a processing time, i.e., the time needed by the machine to process that specific job, and a cost function, which indicates how costly it is for that agent to spend a unit of time in the system. The main objective in sequencing problems consists in finding an optimal order, i.e., an order on the jobs that minimizes the total aggregated cost of all agents. The cost of an agent will naturally depend on its completion time, and this dependence is linear in the classical model (Smith, 1956). There are, however, numerous variants of sequencing problems that allow for an adaptation to real situations. One of them is the sequencing problem with precedence constraints, where some jobs need to be processed before others, as analyzed in Sidney (1975) and Hamers et al. (2005). In most interactive sequencing problems, an initial order is assumed as a starting point and the focus is on the allocation of cost savings with respect to this initial order. For the classical model, Curiel et al. (1989) introduces and axiomatically characterizes an allocation rule, the *equal gain splitting rule* (EGS-rule), based on neighbor switches to derive an optimal order from the initial one, which was later generalized in Hamers et al. (1996).

A GMS-problem is closely related to an MCST problem. However, in a GMS-problem the costs are computed in a different way. In particular, the order in which the edges are activated has a substantial effect on the cost in our setup, whereas this order is irrelevant in calculating the costs in an MCST problem. Besides, the GMS-problem is deeply linked to sequencing problems with precedence constraints. Although interactive sequencing problems with precedence constraints have been treated in the literature before (see Hamers et al., 2005), the approach under which we will study them here has, to the best of our knowledge, never been adopted. In our setting, we start from a graph $(N \cup \{0\}, E)$ where N is the set of nodes corresponding to the agents, 0 is the source node that must serve all agents, and E is a set of edges connecting the nodes. Each edge has a specific activation time, and each agent has a cost depending on the time it gets connected. We will aim, on the one hand, to find an optimal connection order on the agents such that the corresponding total aggregate connection costs over all players are minimized. Note that a connection order on the agents induces an activation order on the edges. On the other hand, we aim to find a fair allocation of these costs.

In this work, we start by motivating the GMS-problem and by formally describing the general problem in detail. The difficulty in obtaining an optimal connection order for the GMS-problem results in the restriction to GMS-problems on trees. We first focus and discuss a procedure to find an optimal order for network structures consisting of 2 lines arising from the source, integrated with an allocation approach. The proposed solution algorithm for these $2-$lines GMS-problems is a reformulation of the work of Sidney (1975), but including a more elaborate procedure and additional ingredients like merge segments that will be essential for our cost allocation procedure. The $2-$lines algorithm and allocation procedure serve as the basis to recursively solve $n-$lines and general tree GMS-problems. For the allocation procedure, we use a myopic reference order that will depend on the problem at hand and show that it is possible to go from the reference order to an optimal order by non-negative savings by switching blocks of agents appropriately selected on the basis of merge segments.

The remainder of this work is organized as follows. Section 2 focuses on the general GMS-problem. Section 3 provides a solution algorithm and an allocation rule for $2-$lines GMS-problems. Sections 4 and 5 generalize the previous procedures to $n-$lines and tree GMS-problems, respectively. Finally, Section 6 presents a discussion and final remarks of this work.

2. PROBLEM DESCRIPTION

In this section we will formally present the problem under consideration, as well as some essential definitions to address it. From now on we will use the terms machine and players instead of source and nodes, respectively.

A *graph machine scheduling problem*, GMS-problem, can be summarized by a tuple $\mathcal{G} = (N, 0, E, \gamma, \alpha)$, where N is a finite set of jobs or players, 0 represents the machine, E is a set of available (precedence) edges between players and machine, i.e., $E \subseteq \{\{i, j\} \mid i, j \in N \cup \{0\}, i \neq j\}$, such that $(N \cup \{0\}, E)$ is a connected graph, $\gamma\colon E \to \mathbb{R}_+$ with $\gamma_{ij} = \gamma(\{i, j\})$ representing the activation time of the edge $\{i, j\} \in E$, and, finally, $\alpha\colon N \to \mathbb{R}_+$, with $\alpha(i)$ representing the linear cost coefficient to spend one time unit in the system for player $i \in N$. The main assumption is that a player $i \in N$ can only be processed by the machine if all players on a path in E from i to the machine have been processed before. A processing or connection order is described by a bijection $\sigma\colon N \to \{1, \ldots, |N|\}$, and $\Pi(N)$ denotes the set of all processing orders.

Definition 1. Let $(N, 0, E, \gamma, \alpha)$ be a GMS-problem, and let $\sigma \in \mathcal{F}(N)$. Given $i \in N$, we define the *completion time of player i with respect to σ*, $C_i(\sigma)$, as follows:

$$C_i(\sigma) = \sum_{k \in P_\sigma(i)} C_k(\sigma) + \min\left\{\gamma_{ij} \mid j \in P^0_\sigma(i) \text{ and } \{i, j\} \in E\right\}.$$

Given $i \in N$, we define the *cost of player i with respect to σ*, $c_i(\sigma)$, as follows:

$$c_i(\sigma) = \alpha(i) \cdot C_i(\sigma).$$

Let $c(\sigma) = (c_i(\sigma))_{i \in N}$ denote the individual cost vector with respect to σ. We define the *total cost of σ*, $TC(\sigma)$, as follows:

$$TC(\sigma) = \sum_{i \in N} c_i(\sigma). \tag{1}$$

Among other things, this work aims to determine an optimal order $\hat{\sigma} \in \mathcal{F}(N)$ that minimizes the total costs among all feasible processing orders. It is important to note that the problem of finding an optimal connection order for general graphs $(N \cup \{0\}, E)$ is hard. In this work, we will focus on GMS-problems such that $(N \cup \{0\}, E)$ is a tree.

It should be stressed that by restricting the problem to trees, there will be only one path between any two nodes. With the purpose of simplifying the notation, the GMS-problems treated from now on will be denoted by a tuple $(N, 0, E, \gamma, \alpha)$, where γ now is a function on N. In particular, for $\gamma\colon N \to \mathbb{R}_+$ we have that $\gamma(i) = \gamma_{ip^0_E(i)}$, where $p^0_E(i)$ is the first player on the unique path between the machine and i. In this way, it is easily seen that equation (1) can be reformulated as

$$TC(\sigma) = \sum_{k=1}^{|N|} \gamma(\sigma^{-1}(k)) \cdot \left(\sum_{\{j \in N \mid \sigma(j) \geq k\}} \alpha(j) \right).$$

3. 2−LINES GMS-PROBLEMS

In this section we describe and analyze 2-lines GMS-problems. We will first tackle the question of how to find an optimal connection order for these problems, and second how to allocate the cost of such a minimal connection order among the players involved.

3.1. Optimal orders

A GMS-problem $(N, 0, E, \gamma, \alpha)$ is called a *2−lines GMS-problem* if there exists a partition $\langle A, B\rangle$ of N with $A = \{a_1, \ldots, a_{\tilde{s}}\}$ and $B = \{b_1, \ldots, b_{\tilde{t}}\}$ with $\tilde{s} + \tilde{t} = |N|$, such that

$$E = \left\{\{0, a_1\}, \{a_1, a_2\}, \ldots, \{a_{\tilde{s}-1}, a_{\tilde{s}}\}\right\} \cup \left\{\{0, b_1\}, \{b_1, b_2\}, \ldots, \{b_{\tilde{t}-1}, b_{\tilde{t}}\}\right\}.$$

The sets A and B are called branches. For this particular case, a feasible order is described by a bijection $\sigma\colon A \cup B \to \{1, 2, \ldots, \tilde{s} + \tilde{t}\}$ such that

$$\sigma(a_k) < \sigma(a_l) \Rightarrow k < l;$$
$$\sigma(b_k) < \sigma(b_l) \Rightarrow k < l.$$

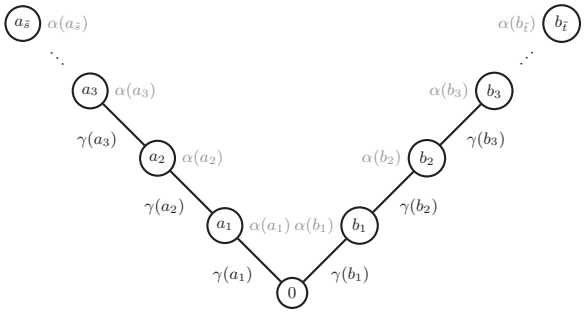


Figure 1: Graphical representation of a 2−lines GMS-problem.

Let $\mathcal{F}(A \cup B)$ denote the set of all such feasible orders. A graphical representation of a 2−lines GMS-problem can be seen in Figure 1.

Definition 2. Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem, and let $\sigma \in \mathcal{F}(A \cup B)$. We define a *segment of A from h to l* as the following subset of A: $Q_{hl} = \{a_h, a_{h+1}, \ldots, a_l\}$, where $1 \leq h \leq l \leq \tilde{s}$. Analogously, we define a *segment of B from h to l* as the following subset of B: $R_{hl} = \{b_h, b_{h+1}, \ldots, b_l\}$, where $1 \leq h \leq l \leq \tilde{t}$.

When we do not specify which branch a certain segment belongs to, we will use the notation X or Y instead of Q and R. A segment from the beginning of a branch is called a *head*.

Definition 3. Given a segment X, we define the *cost weighted average time per edge of X*, $CAT(X)$, as

$$CAT(X) = \frac{\sum_{i \in X} \gamma(i)}{\sum_{i \in X} \alpha(i)}.$$

Moreover, given a segment X, we define its *urgency*, $U(X)$, as $U(X) = \frac{1}{CAT(X)}$.

Given a 2−lines GMS-problem, $(N, 0, E, \gamma, \alpha)$, our objective is to solve the following problem:

$$\begin{aligned} \min \quad & TC(\sigma) \\ \text{s.t.} \quad & \sigma \in \mathcal{F}(A \cup B). \end{aligned}$$

In Algorithm 1 we formally present the algorithm to solve 2−lines GMS-problems. Steps 1–3 constitute an iteration of the algorithm. The output of Algorithm 1 is a *merge order*: $\hat{\sigma} = (M_1, M_2, \ldots, M_{m-1}, M_m)$, with $m \geq 2$, where $M_1, M_2, \ldots, M_m$ are called *merge segments*. It could be possible that both M_k and M_{k+1} belong to the same branch (because they might be merged at different steps). Of course, a merge order $\hat{\sigma}$ corresponds to one order on all players. If we do not want to highlight the merge segments, we will use the notation $\hat{\tau}$ for this order.

Next, we will show a series of results to prove that Algorithm 1 leads to an optimal order.

Remark 1. Given a set $Z \subseteq N$, we will use the notation $\gamma[Z] = \sum_{i \in Z} \gamma(i)$ and $\alpha[Z] = \sum_{i \in Z} \alpha(i)$. Hence, if X is a segment we can write $CAT(X) = \frac{\gamma[X]}{\alpha[X]}$.

The following proposition shows that an optimal order cannot have two consecutive segments of the same branch separated by nodes of the other branch when the CAT of the first segment is greater than the CAT of the next segment.

Proposition 1. *Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem. Let X and Y be two segments that belong to the same branch, and let Z be a segment from the other branch. If $CAT(X) > CAT(Y)$, then $\tau = (\sim, X, Z, Y, \sim)$ is not optimal.*

Definition 4. Let $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$ be the output of Algorithm 1 for a 2−lines GMS-problem and let τ be a feasible order. Let $k \in \{1, 2, \ldots, m\}$. A *component of M_k in τ* is defined as a maximal connected subset of M_k with respect to τ. Denote by M_k/τ the *set of components of M_k in τ*, and let $M_k/\tau = \{G_1, G_2, G_3, \ldots, G_{m_k}\}$ be the different components in the order they appear in τ, that is: $\tau = (\sim, G_1, \ldots, G_2, \ldots, G_3, \ldots, G_{m_k}, \sim)$.

Algorithm 1 Algorithm to solve a 2−lines GMS-problem

0. Initialize $k = 1, i = 1$.
1. Consider the 2−lines GMS-problem $(N, 0, E, \gamma, \alpha)$. Initialize $l = 1, l' = 1$.
   - If $\frac{\gamma(a_1)}{\alpha(a_1)} \leq \frac{\gamma(b_1)}{\alpha(b_1)}$, select pivot $P_i = \{a_1\}$.
   - If $\frac{\gamma(a_1)}{\alpha(a_1)} > \frac{\gamma(b_1)}{\alpha(b_1)}$, select pivot $P_i = \{b_1\}$.
2. (a) If $P_i \subseteq A$, take $l = l + 1$.
   - If $l \leq \tilde{t}$, compare $CAT(P_i)$ to $CAT(R_{1l})$.
     - If $CAT(P_i) \leq CAT(R_{1l})$, go back to step 2.
     - If $CAT(P_i) > CAT(R_{1l})$, then $i = i + 1$, $P_i = R_{1l}$. Go back to step 2.
   - If $l > \tilde{t}$, go to step 3.

   (b) If $P_i \subseteq B$, take $l' = l' + 1$.
   - If $l' \leq \tilde{s}$, compare $CAT(P_i)$ to $CAT(Q_{1l'})$.
     - If $CAT(P_i) \leq CAT(Q_{1l'})$, go back to step 2.
     - If $CAT(P_i) > CAT(Q_{1l'})$, then $i = i + 1$, $P_i = Q_{1l'}$. Go back to step 2.
   - If $l' > \tilde{s}$, go to step 3.
3. Set $M_k = P_i$.

   (a) If $M_k \subseteq A$.
   - If $A \setminus M_k = \emptyset$, then $M_{k+j} = \{b_j\}$ for all $j \in \{1, \ldots, \tilde{t}\}$. The algorithm is finished.
   - Otherwise, let $A = r(A \setminus M_k)$, where $r\colon A \setminus M_k \to A$ is a renumbering function such that $r(a_h) = a_{h-l'}$, for all $h \in \{l' + 1, \ldots, \tilde{s}\}$. Let $\gamma(a_h) = \gamma(a_{h+l'})$ and $\alpha(a_h) = \alpha(a_{h+l'})$ for all $h \in \{1, \ldots, \tilde{s} - l'\}$. Set $N = A \cup B$, $k = k + 1$ and $i = i + 1$. Go back to step 1.

   (b) If $M_k \subseteq B$.
   - If $B \setminus M_k = \emptyset$, then $M_{k+j} = \{a_j\}$ for all $j \in \{1, \ldots, \tilde{s}\}$. The algorithm is finished.
   - Otherwise, let $B = \tilde{r}(B \setminus M_k)$, where $\tilde{r}\colon B \setminus M_k \to B$ is a renumbering function such that $r(b_h) = b_{h-l}$, for all $h \in \{l + 1, \ldots, \tilde{t}\}$. Let $\gamma(b_h) = \gamma(b_{h+l})$ and $\alpha(b_h) = \alpha(b_{h+l})$ for all $h \in \{1, \ldots, \tilde{t} - l\}$. Set $N = A \cup B$, $k = k + 1$ and $i = i + 1$. Go back to step 1.

Given a non-connected merge segment, the following lemma guarantees that there will be at least one pair of consecutive components in a feasible order such that the CAT of the first component is strictly greater than the CAT of the next component.

Lemma 1. *Let $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$ be the output of Algorithm 1 for a 2−lines GMS-problem. Take $k \in \{1, 2, \ldots, m\}$ such that $|M_k| > 1$. And let τ be a feasible order such that $\left|M_k/\tau\right| > 1$. Then, there exists $k \in \{2, 3, \ldots, m_k\}$ such that $CAT(G_{k-1}) > CAT(G_k)$.*

The following lemma tells us that there can be no optimal order in which merge segments have more than one component.

Lemma 2. *Let $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$ be the output of Algorithm 1 for a 2−lines GMS-problem. It holds that the elements of $M_k, k \in \{1, \ldots, m\}$, are consecutive in any optimal order.*

The proposition below states that at least one optimal order has to start with the first merge segment obtained by applying Algorithm 1.

Proposition 2. *Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem. Let $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$ be the output of Algorithm 1 for such problem. There always exists an optimal order that starts with M_1.*

The following proposition shows that the specific structure of an optimal order leads to an optimal order of a subproblem.

Proposition 3. *Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem, and let $\tilde{\sigma}^N = (M_1, \ldots, M_m)$ be the output of Algorithm 1. Let τ^N be an optimal order. If τ^N starts with M_1, it holds that $\tau^N|_{N \setminus M_1}$ is an optimal order for the subproblem on $N \setminus M_1$.*

The next result describes a reverse version of Proposition 3: if we have a specific optimal order for a subproblem, we can derive an optimal order for the general problem.

Lemma 3. *Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem, and let $\tilde{\sigma}^N = (M_1, \ldots, M_m)$ be the output of Algorithm 1. Let $\tau^{N\setminus M_1}$ be an optimal order for the problem on $N \setminus M_1$. It holds that $\tau^N = (M_1, \tau^{N\setminus M_1})$ is an optimal order for $(N, 0, E, \gamma, \alpha)$.*

We present below the main result of this subsection, which indicates that Algorithm 1 always leads to an optimal order.

Theorem 1. *Let $(N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem, and let $\hat{\tau}$ be the order provided by Algorithm 1. Then, $TC(\hat{\tau}) \leq TC(\tau)$ for all $\tau \in \mathcal{F}(A \cup B)$.*

Proof. The proof uses induction to the number of players, $|N|$. Consider $|N| = 2$. In such a case, there are two possible orders, $\tau_1 = (a_1, b_1)$ and $\tau_2 = (b_1, a_1)$. Note that:

$$TC(\tau_1) = (\alpha(a_1) + \alpha(b_1)) \cdot \gamma(a_1) + \alpha(b_1) \cdot \gamma(b_1);$$
$$TC(\tau_2) = (\alpha(b_1) + \alpha(a_1)) \cdot \gamma(b_1) + \alpha(a_1) \cdot \gamma(a_1),$$

and thus

$$TC(\tau_1) - TC(\tau_2) = \alpha(b_1) \cdot \gamma(a_1) - \alpha(a_1) \cdot \gamma(b_1) = \alpha(b_1) \cdot \alpha(a_1) \cdot \left(\frac{\gamma(a_1)}{\alpha(a_1)} - \frac{\gamma(b_1)}{\alpha(b_1)} \right). \quad (2)$$

Algorithm 1 compares $\frac{\gamma(a_1)}{\alpha(a_1)}$ to $\frac{\gamma(b_1)}{\alpha(b_1)}$ in order to choose the first merge segment, which in this case will consist of a single node. From (2), we can see that the optimal order will be determined by the exact same comparison, thus Algorithm 1 leads to an optimal order.

Now assume that Algorithm 1 leads to an optimal order if the number of players is $k < |N|$.

Now, take $k = |N|$. Let $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$ be the output of Algorithm 1 corresponding to $\hat{\tau}$. Naturally, $\hat{\sigma}|_{N\setminus M_1} = (M_2, \ldots, M_m)$ will be an output of our procedure for the problem with set of players $N \setminus M_1$. Using our induction hypothesis, $\hat{\sigma}|_{N\setminus M_1}$ is optimal for such subproblem. From Lemma 3, the order $(M_1, \hat{\sigma}|_{N\setminus M_1})$ is optimal. Clearly, $\hat{\sigma} = (M_1, \hat{\sigma}|_{N\setminus M_1})$, finishing the proof. □

3.2. Allocating the minimal cost

This subsection introduces the κ rule as a cost allocation rule for 2−lines GMS-problems. The κ rule takes as a reference point a myopic connection order and its corresponding cost vector and will subtract a specific allocation vector of the cost savings as given by the block splitting rule (BSR), which will be described later. The underlying allocation procedure is closely tied to the theoretical results presented in Subsection 3.1. In particular, the merge segments will be the foundation of the allocation procedure that we will discuss. Below, we explain the ideas behind the κ rule in more detail.

Let $\mathcal{G} = (N, 0, E, \gamma, \alpha)$ be a 2−lines GMS-problem. The order that we will use as a reference point is an endogenous and myopic order τ_0. It will depend on the particular problem we are considering, in the following way: at each step, the machine selects the player that has a higher urgency, always taking into account the existing precedence relations. For expositional simplicity, we will assume that $\frac{\alpha(i)}{\gamma(i)} \neq \frac{\alpha(j)}{\gamma(j)}$ for all $i, j \in N, i \neq j$, i.e., all players' urgencies are different. Thus, τ_0 is unique[1]. Moreover, given an optimal order $\hat{\tau}$, the total amount that will be saved is $g^N = TC(\tau_0) - TC(\hat{\tau})$. The allocation approach starts from the reference order, τ_0, and "repairs" it until the optimal order found by Algorithm 1, $\hat{\tau}$, is reached. In order to guarantee that these repairs lead to non-negative cost savings and there is a local incentive to perform each step, we exchange blocks of players related to the merge segments. But how do we choose these blocks in general in a unique way such that non-negative switching gains are guaranteed in each step? The determination of these blocks cannot be carried out simply by observing the orders τ_0 and $\hat{\tau}$, but will be done iteratively. To do that, we will present in detail the *block splitting rule* (BSR). The key point of this approach is to determine, at each step, which blocks are to be swapped. Note that given two consecutive blocks, X and Y with X before Y, the gain resulted from switching them is:

$$g_{XY} = \alpha[Y] \cdot \gamma[X] - \alpha[X] \cdot \gamma[Y] = \alpha[Y] \cdot \alpha[X] \cdot \frac{\gamma[X]}{\alpha[X]} - \alpha[X] \cdot \alpha[Y] \cdot \frac{\gamma[Y]}{\alpha[Y]}$$
$$= \alpha[Y] \cdot \alpha[X] \cdot (CAT(X) - CAT(Y)).$$

[1] In case of ties, any possible reference order is considered with a certain probability. We will elaborate on this issue in Section 6.

The merge segments play a fundamental role in defining the BSR, since by conveniently using their properties along with Algorithm 1 we will be able to guarantee non-negative savings at each iteration. Thus, this procedure consists of two main stages: firstly, we will repair those merge segments whose players are not consecutive, and secondly reorder them as in $\hat{\tau}$. To this end, we will need to consider $\hat{\sigma}$, the corresponding merge order to $\hat{\tau}$. Algorithm 2 shows the scheme of this procedure.

Algorithm 2 Algorithm to allocate the gains of a 2−lines GMS-problem

0. Obtain τ_0 and apply Algorithm 1 to get $\hat{\sigma} = (M_1, M_2, \ldots, M_m)$. Initialize $k = 1$, $r = 1$, it $= 1$, and $\tau' = \tau_0$.
1. (a) If $|M_k/\tau'| = 1$, take $k = k + 1$.
   - If $k \leq m - 1$, go back to step 1.
   - If $k = m$, go to step 2.

   (b) If $|M_k/\tau'| > 1$, take $\tilde{k} \in \{2, \ldots, m_k\}$ such that $CAT(G_{\tilde{k}-1}) > CAT(G_{\tilde{k}})$ (we know there exists such a pair of components from Lemma 1). It is clear that between $G_{\tilde{k}-1}$ and $G_{\tilde{k}}$ there are only players from the other branch, i.e.,
   $$\tau' = (\sim, G_{\tilde{k}-1}, Z, G_{\tilde{k}}, \sim),$$
   where Z is a segment from the opposite branch of M_k. From Proposition 1, we know that either the order $\tau_1 = (\sim, G_{\tilde{k}-1}, G_{\tilde{k}}, Z, \sim)$ or the order $\tau_2 = (\sim, Z, G_{\tilde{k}-1}, G_{\tilde{k}}, \sim)$ has a lower total cost than τ'. Take $\tau'' = \arg\min_{\tau}\{TC(\tau) \colon \tau \in \{\tau_1, \tau_2\}\}$.
   - If $\tau'' = \tau_1$, then the blocks that have been switched are Z and $G_{\tilde{k}}$. Players from Z should receive $\frac{1}{2|Z|} g_{ZG_{\tilde{k}}}$, while players from $G_{\tilde{k}}$ should receive $\frac{1}{2|G_{\tilde{k}}|} g_{ZG_{\tilde{k}}}$.
   - If $\tau'' = \tau_2$, then the blocks that have been switched are $G_{\tilde{k}-1}$ and Z. Players from $G_{\tilde{k}-1}$ should receive $\frac{1}{2|G_{\tilde{k}-1}|} g_{G_{\tilde{k}-1}Z}$, while players from Z should receive $\frac{1}{2|Z|} g_{G_{\tilde{k}-1}Z}$.

   Set $\tau' = \tau''$, and take it $=$ it $+ 1$. Go back to step 1.
2. (a) If $r \leq m$, consider the bijection
   $$\begin{aligned} \rho \colon \{1, 2, \ldots, m\} &\to \{1, 2, \ldots, m\} \\ i &\mapsto \rho(i) = j, \end{aligned}$$
   such that $\tau' = (M_{\rho(1)}, \ldots, M_{\rho(m)})$. We need to go from τ' to $\hat{\tau}$.
   i. If $\rho(r) = r$, take $r = r + 1$. Go back to step 2.
   ii. If $\rho(r) \neq r$, take $\tilde{r} \in \{r + 1, \ldots, m\}$ such that $\rho(\tilde{r}) = r$ (this means that M_r is on position $\tilde{r}$, i.e., $M_r = M_{\rho(\tilde{r})}$). By Algorithm 1, it holds that
   $$CAT(M_r) \leq CAT\left(M_r^{\cup}\right),$$
   where $M_r^{\cup} = \bigcup_{l=r}^{\tilde{r}-1} M_{\rho(l)}$. Hence, the order
   $$\tau'' = (\sim, M_{\rho(\tilde{r})}, M_{\rho(r)}, \ldots, M_{\rho(\tilde{r}-1)}, M_{\rho(\tilde{r}+1)}, \sim)$$
   that consists in moving $M_{\rho(\tilde{r})}$ to the front of $M_{\rho(r)}$ (so that $M_{\rho(\tilde{r})} \equiv M_r$ is now on position r) has lower total cost than τ'. The blocks that have been switched are $M_r^{\cup}$ and M_r. Allocate $\frac{1}{2|M_r^{\cup}|} g_{M_r^{\cup} M_r}$ to the players in $M_r^{\cup}$ and $\frac{1}{2|M_r|} g_{M_r^{\cup} M_r}$ to the players in M_r. Set $\tau' = \tau''$, and take $r = r + 1$ and it $=$ it $+ 1$. Go back to step 2.

   (b) If $r > m$, then $\tau' = \hat{\tau}$ and $I =$ it $- 1$. The algorithm is finished. The outputs of the algorithm are the allocation and the misplaced blocks at each iteration it $\in \{1, \ldots, I\}$.

Let X^{it} and Y^{it} be the misplaced blocks switched at iteration "it" of Algorithm 2. We define:

$$\text{BSR}^{\text{it}}(\tau_0, \hat{\sigma}) = \frac{1}{2} g_{X^{\text{it}} Y^{\text{it}}} \left(\frac{1}{|X^{\text{it}}|} e^{X^{\text{it}}} + \frac{1}{|Y^{\text{it}}|} e^{Y^{\text{it}}} \right),$$

where, for $S \subseteq N$, e^S is the vector in $\mathbb{R}^N$ satisfying $e_i^S = 1$ if $i \in S$ and $e_i^S = 0$ otherwise. $\text{BSR}^{\text{it}}(\tau_0, \hat{\sigma})$ represents the allocation obtained at iteration it, hence $\text{BSR}(\tau_0, \hat{\sigma}) = \sum_{\text{it}=1}^{I} \text{BSR}^{\text{it}}(\tau_0, \hat{\sigma})$, where I represents the total number of iterations needed. Subsequently, we define the cost allocation rule, κ, by setting

$$\kappa(\mathcal{G}) = c(\tau_0) - \text{BSR}(\tau_0, \hat{\sigma}),$$

for a 2−lines GMS-problem $\mathcal{G}$.

4. $n-$LINES GMS-PROBLEMS

This section generalizes the optimization and allocation results for the 2−lines GMS-problems to $n-$lines GMS-problems.

4.1. Optimal orders

A GMS-problem $(N, 0, E, \gamma, \alpha)$ is called an *$n-$lines GMS-problem* if there exists a partition $\langle A^1, \dots, A^n \rangle$ of N with $A^k = \{a_1^k, \dots, a_{\tilde{s}_k}^k\}$ for all $k \in \{1, \dots, n\}$ with $\sum_{k=1}^n \tilde{s}_k = |N|$, such that

$$E = \bigcup_{k=1}^{n} \big\{\{0, a_1^k\}, \{a_1^k, a_2^k\}, \dots, \{a_{\tilde{s}_k - 1}^k, a_{\tilde{s}_k}^k\}\big\}.$$

As for the 2−lines GMS-problems, the sets $A^k, k \in \{1, \dots, n\}$, are called branches. A feasible order is described by a bijection $\sigma \colon N \to \{1, 2, \dots, |N|\}$ such that $\sigma(a_h^k) < \sigma(a_l^k) \Rightarrow h < l$, for all $k \in \{1, \dots, n\}$. Let $\mathcal{F}(N)$ denote the set of all such feasible orders. The definitions regarding the segments and $CATs$ can be directly extended to this generalization. A graphical representation of an $n-$lines GMS-problem is provided in Figure 2.

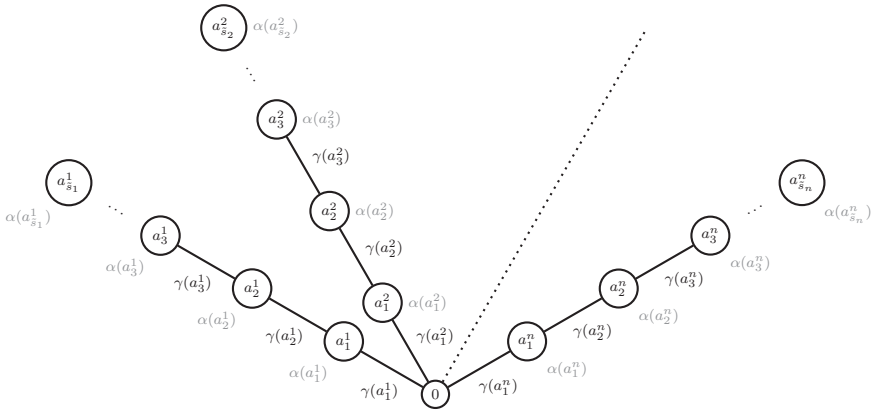


Figure 2: Graphical representation of an $n-$lines GMS-problem.

To solve an $n-$lines GMS-problem, we propose an algorithm that combines the concept of recursion with Algorithm 1. This procedure is based on the following idea: given an $n-$lines GMS-problem $(N, 0, E, \gamma, \alpha)$, we consider a 2−lines GMS-problem, $(A^h \cup A^l, 0, E|_{A^h \cup A^l}, \gamma, \alpha)$, where $h, l \in \{1, \dots, n\}$. Note that $E|_{A^h \cup A^l}$ is the restriction of E to the players of $A^h \cup A^l$. $\hat{\sigma}_{hl}$ will denote the output of Algorithm 1 for such a subproblem in which the merge segments are specified. $\hat{\tau}_{hl}$ will refer to the corresponding order on all players involved. Furthermore, $A^{hl}_{\hat{\tau}}$ will denote the branch formed by the nodes from A^h and A^l following the order specified by $\hat{\tau}_{hl}$. We informally present the algorithm below.

Algorithm 3 Algorithm to solve an $n-$lines GMS-problem

1. Consider an $n-$lines GMS-problem $(N, 0, E, \gamma, \alpha)$. Select branches A^1 and A^2.
2. Apply Algorithm 1 to solve the corresponding 2−lines GMS-problem, $(A^1 \cup A^2, 0, E|_{A^1 \cup A^2}, \gamma, \alpha)$. This leads to an optimal order, $\hat{\tau}_{12}$. Replace A^1 and A^2 with the branch $A^{12}_{\hat{\tau}}$. We get a new problem with one branch less. Renumber the branches adequately.
3. (a) If there are still more than two branches left, go back to step 1.
   (b) If there are two branches left, apply Algorithm 1. The order obtained is the solution.

Along similar lines as in Lemma 2 one can show the following result.

Lemma 4. *Let $(N, 0, E, \gamma, \alpha)$ be an $n-$lines GMS-problem, and $\hat{\sigma}_{12} = (M_1, M_2, \dots, M_m)$ be the output of Algorithm 1 for the 2−lines GMS-problem $(A^1 \cup A^2, 0, E|_{A^1 \cup A^2}, \gamma, \alpha)$. It holds that the elements of $M_k, k \in \{1, \dots, m\}$, are consecutive in any optimal order for $(N, 0, E, \gamma, \alpha)$.*

Moreover, the result below guarantees that obtaining an optimal order for an $n-$lines GMS-problem, the first two branches can be replaced by one branch that reflects the optimal order found by Algorithm 1 for the corresponding 2−lines subproblem.

Proposition 4. *Let $(N, 0, E, \gamma, \alpha)$ be an $n-$lines GMS-problem, and let τ_{12}^* be the output of Algorithm 1 for the 2−lines GMS-problem $(A^1 \cup A^2, 0, E|_{A^1 \cup A^2}, \gamma, \alpha)$. There exists an optimal order $\hat{\tau}$ for $(N, 0, E, \gamma, \alpha)$ such that $\hat{\tau}(i) < \hat{\tau}(j)$ for all $i, j \in A^1 \cup A^2$ for which $\tau_{12}^*(i) < \tau_{12}^*(j)$.*

The following result states that Algorithm 3 leads to an optimal order, and can be proved by induction in the number of branches.

Theorem 2. *Let $(N, 0, E, \gamma, \alpha)$ be an $n-$lines GMS-problem, and let $\hat{\tau}$ be an order provided by Algorithm 3. Then, $TC(\hat{\tau}) \leq TC(\tau)$ for all $\tau \in \mathcal{F}(N)$.*

4.2. Allocating the minimal cost

We will illustrate how to extend the ideas of the allocation procedure as described by κ for 2−lines GMS-problems into a rule on $n-$lines GMS-problems. Consider an $n-$lines GMS-problem $\mathcal{G} = (N, 0, E, \gamma, \alpha)$. Again, take as starting point of the allocation mechanism a reference order on all players, τ_0^N, in the same way as before. Let $\hat{\tau}^N$ be the optimal order provided by Algorithm 3. The total savings of $g^N = TC(\tau_0^N) - TC(\hat{\tau}^N)$ need to be adequately subtracted from $c(\tau_0^N)$ to obtain the final cost allocation. To determine the proportion of g^N for which each player is responsible, we apply a procedure similar to that in Algorithm 3: we will recursively allocate the local savings obtained at the 2−lines GMS-problems that comprise our $n-$lines. The sum of these local savings however is not necessarily equal to g^N. Instead, we use these numbers to determine the *relative* importance of the different subproblems (and, mainly, of the players involved in these problems) in the final savings obtained.

Let $\tau_0^k, k \in \{2, \ldots, n\}$ be the reference order for the 2−lines GMS-problem induced by branches $A^{1\ldots k-1}$ and A^k, and let $\hat{\tau}^k$ and $\hat{\sigma}^k$ be the optimal order and its corresponding merge order provided by Algorithm 1 for such 2−lines GMS-problem. The local cost savings are defined by $g^k = TC(\tau_0^k) - TC(\hat{\tau}^k)$ and they will be allocated in the same way as before for each subproblem. That is, given $k \in \{2, \ldots, n\}$, going from τ_0^k to $\hat{\sigma}^k$ leads to a saving of g^k that is allocated among the players involved using Algorithm 2, thus obtaining $\mathrm{BSR}(\tau_0^k, \hat{\sigma}^k)$. Each of these vectors are now complemented with 0's on those coordinates that refer to non-involved players. Figure 3 summarizes this procedure.

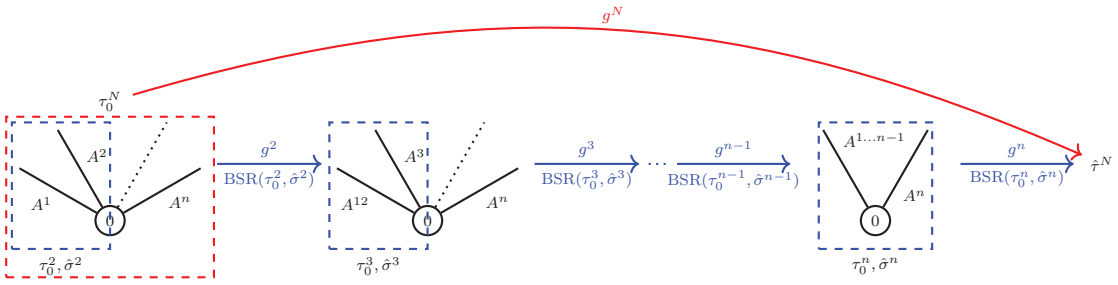


Figure 3: Steps of the allocation procedure in an $n-$lines GMS-problem.

We define the cost allocation rule, κ, by setting:

$$\kappa(\mathcal{G}) = c(\tau_0^N) - \frac{g^N}{\sum_{k=2}^n g^k} \cdot \sum_{k=2}^n \mathrm{BSR}(\tau_0^k, \hat{\sigma}^k),$$

for an $n-$lines GMS-problem $\mathcal{G}$.

5. TREE GMS-PROBLEMS

We will now extend the results we have seen for n-lines GMS-problems to the case of tree GMS-problems, both for optimization and cost allocation.

5.1. Optimal orders

A GMS-problem $(N, 0, E, \gamma, \alpha)$ is called a *tree GMS-problem* if $(N \cup \{0\}, E)$ is a tree.

Definition 5. Let $(N, 0, E, \gamma, \alpha)$ be a tree GMS-problem. We define the *degree of a node* $a \in N$, $\deg(a)$, as the number of edges incident on that node.

Definition 6. Let $(N, 0, E, \gamma, \alpha)$ be a tree GMS-problem. A *sub-source* will be either the machine, 0, or a node with degree at least 3. Let $\mathcal{S}$ be the set of sub-sources.

Given the sub-sources of a tree, we are interested in knowing their *level*.

Definition 7. Let $(N, 0, E, \gamma, \alpha)$ be a tree GMS-problem. The *level* $\ell(s)$ *of a sub-source* $s \in \mathcal{S}$ is the number of sub-sources in the path between 0 and s, including 0. Thus, the machine 0 is the only sub-source with level 1.

We assume an ordering on the sub-sources, from level 1 to the highest level, v. Given a level $l \in \{2, \ldots, v\}$, there are m_l sub-sources. Thus, we can write

$$\mathcal{S} = \{0, s_1^2, \ldots, s_{m_2}^2, s_1^3, \ldots, s_{m_3}^3, \ldots, s_1^v, \ldots, s_{m_v}^v\},$$

where s_k^l denotes the $k-$th sub-source from level l. Figure 4 provides an illustration of the sub-sources of a tree and their levels.

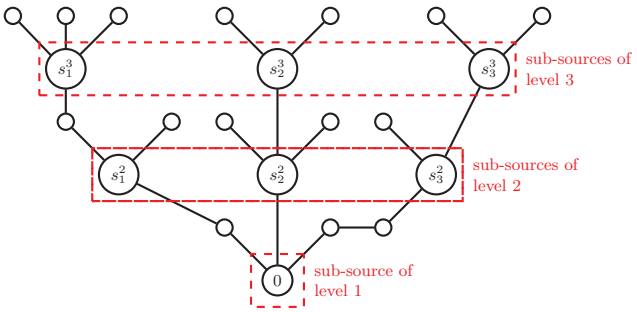


Figure 4: Sketch of the sub-sources of a tree and their levels.

The theoretical results for $n-$lines GMS-problems can be extended to the general case of trees. In particular, given a tree GMS-problem, the elements of the merge segments obtained when solving a $2-$lines GMS-problem of the highest level remain consecutive in any optimal order. Furthermore, there always exists an optimal order that maintains the order induced by the aforementioned subproblem. With all these ingredients, it is immediate to prove that Algorithm 4 below leads to an optimal order. Here, a recursive methodology is adopted, starting with the $n-$lines GMS-problems at the highest level. For each of them, the $2-$lines GMS-problems that comprise it are solved recursively until these $n-$lines are converted into a single line, thus reducing the dimension.

Without proof we state the following result.

Theorem 3. *Let $(N, 0, E, \gamma, \alpha)$ be a tree GMS-problem, and let $\hat{\tau}$ be an order provided by Algorithm 4. Thus, $TC(\hat{\tau}) \leq TC(\tau)$ for all feasible order τ.*

5.2. Allocating the minimal cost

To appropiately extend the κ rule to the context of tree GMS-problems, we need to contemplate the local savings generated in each sub-source, in a similar way to how we made the transition from $2-$lines to $n-$lines GMS-problems.

Algorithm 4 Algorithm to solve a tree GMS-problem

1. Consider the tree GMS-problem $(N, 0, E, \gamma, \alpha)$.
2. Let $k \in \{1, \ldots, m_v\}$. Consider the n_k^v-lines GMS-problem arising from s_k^v. Apply Algorithm 3 to obtain $\tau_{s_k^v}$. Replace the n_k^v branches arising from s_k^v with one branch using the order of $\tau_{s_k^v}$. We have a new tree GMS-problem with one level less since all the subproblems of the highest level have been converted into lines. Renumber the nodes adequately. The highest level has now been reduced by 1.
3. (a) If $\mathcal{S} \neq \{0\}$, go back to step 1.
   (b) If $\mathcal{S} = \{0\}$, solve the resulting $n-$lines GMS-problem with Algorithm 3. The order obtained is the solution.

Let $\mathcal{G} = (N, 0, E, \gamma, \alpha)$ be a tree GMS-problem. For $s \in \mathcal{S}$, we define $N_s = F(s) \cup \{s\}$, where $F(s)$ is the set of followers of s with respect to 0 in the graph $(N \cup \{0\}, E)$. For every sub-source $s \in \mathcal{S}$, we consider an induced $n-$lines GMS-problem on N_s, $(N_s, 0, E|_{N_s}, \gamma, \alpha)$, where all initial branches with respect to s in E have been recursively replaced by a line that corresponds to an optimal order with respect to this branch. Naturally, if $\ell(s) = v$, then $(N_s, 0, E|_{N_s}, \gamma, \alpha)$ is already an $n-$lines GMS-problem and we call it a subproblem at the highest level. Also, given $s \in \mathcal{S}$, let $\tau_0^{N_s}$ and $\hat{\tau}^{N_s}$ denote the corresponding reference order and the optimal order provided by Algorithm 3 for $(N_s, 0, E|_{N_s}, \gamma, \alpha)$, respectively. Subsequently, the *stand-alone cost savings* $w(s)$ *with respect to* s are defined by:

$$w(s) = TC(\tau_0^{N_s}) - TC(\hat{\tau}^{N_s}).$$

Thus, $w(s)$ can be interpreted as local savings made at sub-sources. It should be noted that the sum $\sum_{s\in\mathcal{S}} w(s)$ is not necessarily equal to the total savings $g^N = TC(\tau_0^N) - TC(\hat{\tau}^N)$. As done for $n-$lines GMS-problems, these stand-alone savings help in determining the relative importance of each sub-source to realize g^N.

We will follow a recursive procedure, by first solving the subproblems of the highest level. Once these problems have been solved, the highest level of the tree GMS-problem has been reduced by 1, and we repeat the process. Hence, we will always start from an $n-$lines GMS-problem, which will depend on the specific sub-source we are considering. This is reflected in the notation by writing $n(s), g^k(s), \tau_0^k(s)$, and $\hat{\sigma}^k(s)$. We define the cost allocation rule, κ, by setting:

$$\kappa(\mathcal{G}) = c(\tau_0^N) - g^N \cdot \sum_{s\in\mathcal{S}} \frac{w(s)}{\sum_{t\in\mathcal{S}} w(t)} \cdot \frac{1}{\sum_{k=2}^{n(s)} g^k(s)} \cdot \sum_{k=2}^{n(s)} \mathrm{BSR}(\tau_0^k(s), \hat{\sigma}^k(s)), \tag{3}$$

for a tree GMS-problem $\mathcal{G}$.

Figure 5 displays a flowchart of the allocation procedure for tree GMS-problems.

6. DISCUSSION AND FINAL REMARKS

The procedure presented for calculating the κ rule for $2-$lines, $n-$lines, and tree GMS-problems was limited to the case where all players have different urgencies and, consequently, there is a single reference order (also for all local problems). In the following, we will give some general indications on how to proceed when there are ties. First, we consider all possible reference orders. We assume that, at each step, the machine chooses with equal probability between all jobs with highest urgency. Thus, given the set of all possible reference orders, we will also have a probability distribution on this set. Now, given a fixed reference order, we proceed in the same way as explained in Subsections 3.2, 4.2, and 5.2 obtaining a specific allocation proposal (which will now depend on that reference order). The only difference is that, when solving the $n-$lines GMS-problems associated with each sub-source, the (local) reference order in these subproblems will not be recalculated according to possible ties that may exist. Instead, we will take the restriction of the (general) reference order that we are considering to the players involved in that subproblem. Naturally, the final cost allocation vector is defined as the weighted average of all the reference specific allocation proposals, where the weights are the probabilities of each possible reference order.

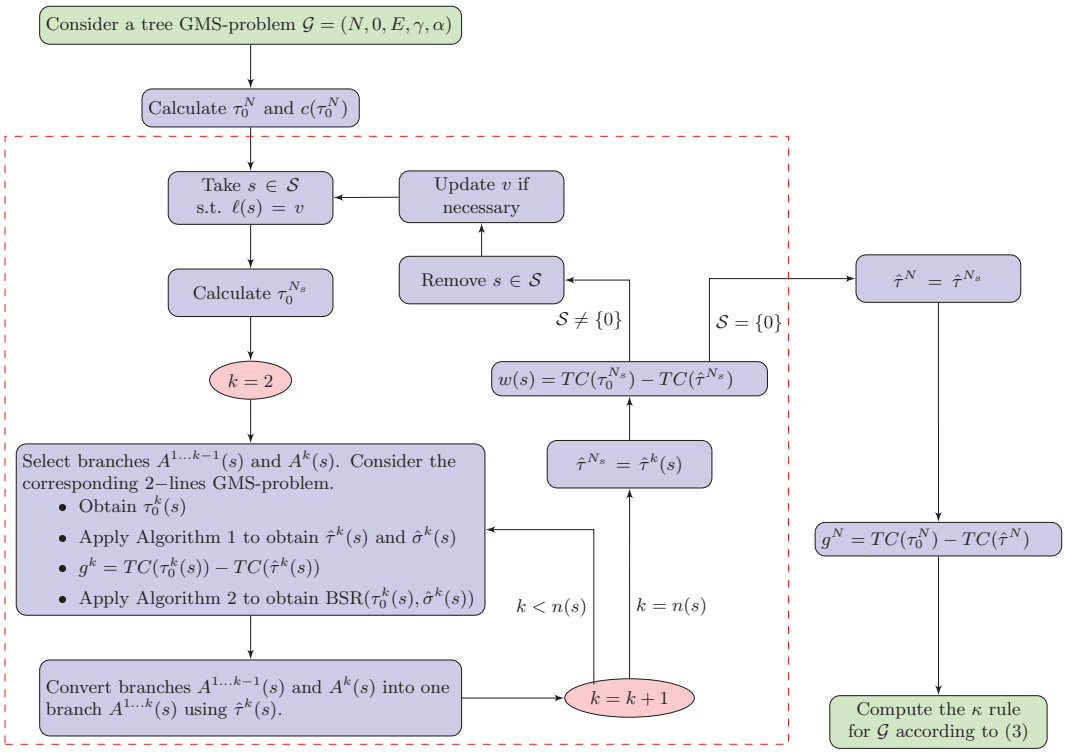


Figure 5: Flowchart of the allocation procedure for tree GMS-problems. The dashed red box contains the steps of Algorithm 3.

As future work, it would be of special interest to find characterizing properties that the proposed allocation rule satisfies. Another open direction of research is to study the allocation aspect of a GMS-problem more directly on the basis of an adequately defined cooperative GMS-game.

REFERENCES

Bergantiños, G., Gómez-Rúa, M., Llorca, N., Pulido, M., Sánchez-Soriano, J. (2014). A new rule for source connection problems. European Journal of Operational Research, 234, 780–788.

Bergantiños, G., Lorenzo, L. (2004). A non-cooperative approach to the cost spanning tree problem. Mathematical Methods of Operations Research, 59, 393–403.

Bergantiños, G., Vidal-Puga, J. (2021). A review of cooperative rules and their associated algorithms for minimum-cost spanning tree problems. SERIEs, 12, 73–100.

Claus, A., Kleitman, D.J. (1973). Cost allocation for a spanning tree. Networks, 3, 289–304.

Curiel, I. (1997). Minimum cost spanning tree games. In Cooperative Game Theory and Applications (pp. 129–148). Springer.

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

Hamers, H., Klijn, F., van Velzen, B. (2005). On the convexity of precedence sequencing games. Annals of Operations Research, 137, 161–175.

Hamers, H., Suijs, J., Tijs, S., Borm, P. (1996). The split core for sequencing games. Games and Economic Behavior, 15, 165–176.

Sidney, J.B. (1975). Decomposition algorithms for single-machine sequencing with precedence relations and deferral costs. Operations Research, 23, 2, 283–298.

Smith, W.E. (1956). Various optimizers for single-stage production. Naval Research Logistics Quarterly, 3, 59–66.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

The least square prenucleolus for games with externalities

Alejandro Saavedra-Nieves[1] 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

In this talk we address the problem of determining the least square prenucleolus for games with externalities (or games in partition function form). Specifically, this solution is based on the minimization of the variance of the excess vectors that can be associated to any allocation vector over the set of all embedded coalitions using a least square criterion. First, an axiomatization of this solution is provided. From a theoretical point of view, we analyse its relations with other proposals of solution for TU games in the game-theoretical literature.

Keywords: games with externalities, least square prenucleolus, optimization problem, excess.

REFERENCES

Ruiz, L. M., Valenciano, F., and Zarzuelo, J. M. (1996). The least square prenucleolus and the least square nucleolus. Two values for TU games based on the excess vector. International Journal of Game Theory, 25, 113–134.

Ruiz, L. M., Valenciano, F., and Zarzuelo, J. M. (1998). The family of least square values for transferable utility games. Games and Economic Behavior, 24, 109–130.

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

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

MEASURES OF RELEVANCE TO THE SUCCESS OF STREAMING PLATFORMS

Juan Carlos Gonçalves-Dosantos[1], Ricardo Martínez[2] and Joaquín Sánchez-Soriano[3]

[1]Universidade da Coruña, Spain.
[2]Universidad de Granada, Spain.
[3]Universidad Miguel Hernández de Elche, Spain.

ABSTRACT

Digital streaming platforms (like Twitch, Spotify, Netflix, Disney, or Kindle) have emerged as one of the main sources of entertainment with a huge potential growth. Many of these platforms distribute royalties among streamers, artists, producers, or writers based on their impact. In this talk we measure the relevance of each of those in the overall success of the platform, an information that may be essential for the allocation of the revenue. We do an axiomatic analysis to provide normative foundations of three relevance metrics: the uniform, the proportional and the subscriber-proportional indicators. The last two indicators implement the so-called pro-rata and user-centric models, which are extensively applied to distribute revenues in the music streaming market. The axioms we propose formalize different principles of fairness, stability and non-manipulability, and are tailor-made for the streaming context. We complete our analysis with a case study, in which we measure the influence of the 19 most worldwide followed streamers on the Twitch platform.

Keywords: Measure, relevance, proportionality, streaming, axiom.

REFERENCES

Alaei, S., Makhdoumi, A., Malekian, A., and Pekeč, S. (2022). Revenue-sharing allocation strategies for two-sided media platforms: pro-rata vs. user-centric. Management Science, 68, 8699–8721.
Adams, W. J., and Yellen, J. L. (1976). Commodity bundling and the burden of monopoly. The Quarterly Journal of Economics, 90, 475–497.
Bergantiños, G., and Moreno-Ternero, J. D. (2023). An axiomatic approach to the revenue sharing in the streaming industry. 22nd annual SAET Conference Paris, France.
Manshadi, V., Niazadeh, R., and Rodilitz, S. (2023). Fair dynamic rationing. Management Science, forthcoming.
Singal, R., Besbes, O., Desir, A., Goyal, V., and Iyengar, G. (2022). Shapley meets uniform: An axiomatic framework for attribution in online advertising. Management Science, 68, 7457–7479.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

A LINEAR MODEL FOR FREIGHT TRANSPORTATION

Alfredo Valencia-Toledo[1] and Juan Vidal-Puga[2]

[1]Universidad Nacional de San Antonio Abad del Cusco, Peru
[2]ECOSOT (Economics, Society and Territory) Universidade de Vigo, Spain

ABSTRACT

We study conflicts generated by mining freight transportation on a linear route that crosses local communities that suffer external costs such as air, water and land pollution. We propose a solution based on stability and fairness principles. In particular, we define a coalitional game with a transferable utility to model the mining freight transportation problem. We present some reasonable properties to characterize an assignment rule to help the planner distribute the compensation among local communities.

Keywords: Socio-economic management, cooperative game theory, core, mining, freight transportation.

1. INTRODUCTION

Mining makes an essential contribution to countries' economies, specially in developing countries, where mining projects are increasingly expected to deliver sustainable benefits to local, regional, and national stakeholders (Gustafsson and Scurrah, 2019). Sustainable development needs to guarantee the investment of the mining industry in the countries' markets and to maintain a peaceful atmosphere in the local communities. A key aspect is to set how to share compensation from the benefit, which is generated by the mining industry. Also, this situation can be found in different contexts like wood, fruit, oil, cement and other industries.

There exist studies based in linear routes, for instance, Ye et al. (2020) optimize transportation system, Lin et al. (2021) study a railway freight transport system, and studies in networks, for instance, Liu et al. (2020) minimize fuel consumption in equilibrated networks, Ma et al. (2018) study urban road networks. Another studies are based on the transports of freight by rivers (Alcalde-Unzu el at., 2021; van den Brink et al., 2018, Sun et al, 2019). Our study lies on the intersection between these two approaches, since our model is lineal and there are networks as alternative routes.

The biggest issue with mineral transportation is to define how far the project area of influence extends when the mineral destination is several hundreds of kilometers away from the mine site, and how inclusive the citizen consultations should be with all the populations involved. Situations that can be modeled in this setting are, for example, a particular case of Peru, one of Latin America's countries with the highest level of mining investment, growth, and economic stability. At the same time, there are many social and environmental conflicts. In particular, the Minerals and Metals Group (MMG) *Las Bambas* transports ore through the Peruvian Southern mining corridor, which is defined as the national road from *MMG Las Bambas* exploitation place (Fuerabamaba, Apurimac) and the train station (Pillones Station, Arequipa). This National road goes through several peasant communities, some of them are considered a direct influence (close to the exploitation place), while others an indirect influence (situated through the road). In this study, we focus only on the indirect influence communities.

Since *MMG Las Bambas* began with the exploitation, social conflicts with the communities have taken place many times in the Apurimac-Cusco region. Some social conflicts have had serious results like life losses. According to the communities, the essential issue is the failure of commitment of the mining firm and the government. According to the mining firm, the essential issue is that communities do not have a clear idea to maintain a solid organization to get their claims.

The mining company has the concession of miner exploitation and it can use the national road to transport the mineral, while local communities complain about all the impact produced by the mineral transportation.

Out of all these issues, we focus on one key question: How can communities be compensated for the transportation of minerals through their land in a way that assures harmony and stability? To answer this question, our model considers a set of players. One of them (the mining firm) is asymmetric with respect to the rest of the players. This player requires the use of the main road. The rest of the players are the local communities. We assume that communities have the option to block the main road to get as much benefit as possible.

Our model fits many real context situations and we study the characteristics and interaction of players, where they try to find the best they can and the asymmetric player (mining firm) has social, economic and environmental commitments to keep harmony in the society. Moreover, this study is related to the social and economic context, where the mining firm has enormous influence, which derives direct and indirect job opportunities for local citizens.

2. COOPERATIVE GAMES

Let $\mathcal{U} = \{1, 2, \dots, \}$ be the universe of (potential) players, and let $\mathcal{N}$ be the set the nonempty, finite subsets of $\mathcal{U}$. A *cooperative game with transferable utility (TU-game)* is a pair (N, v) where $N \in \mathcal{N}$ is a set of players and $v : 2^N \to \mathbb{R}$ is a characteristic function with $v(\emptyset) = 0$, where $v(S)$ is the worth of coalition $S \subseteq N$, which can be interpreted as the benefit that players in S can generate by themselves.

A TU-game (N, v) is *superadditive* if $v(S \cup T) \geq v(S) + v(T)$ for all $S, T \subset N$ with $S \cap T = \emptyset$. This means that two different coalitions can get at least as much benefit working together as separately. A TU-game (N, v) is *monotonic* if $v(S) \leq v(T)$ for all $S \subseteq T \subseteq N$.

For notational convenience, given $y \in \mathbb{R}^S$, we write

$$y(S) = \sum_{i \in S} y_i.$$

An *imputation* of (N, v) is an allocation $x \in \mathbb{R}^N$, satisfying $x(N) = v(N)$ and $x_i \geq v(\{i\})$ for all $i \in N$. We denote the set of imputations as $I(N, v)$, i.e.,

$$I(N, v) = \left\{ x \in \mathbb{R}^N : x(N) = v(N), x_i \geq v(\{i\}) \ \forall i \in N \right\}.$$

The notion of imputation comprises the most basic requirements for a reasonable allocation. It requires that each player receives at least its own stand-alone value. It also requires that the worth of the grand coalition is fully shared, which makes sense under the reasonable condition of superadditivy. Under superadditivy, the set of imputations is always nonempty.

The *core* (Gillies, 1959) of (N, v) is the set of stable imputations, defined as:

$$Core(N, v) = \{ x \in I(N, v) : x(S) \geq v(S) \ \ \forall S \subset N \}.$$

The core has also an intuitive interpretation. We are interested in payoff allocations where no coalition of players can improve by themselves. Also, the core implies the efficiency and individual rationality of players. The main problem with the core is that it may be empty even for superadditive games.

An *assigning rule* is a function that assigns to each TU-game (N, v) in a class of games a vector $\phi(N, v) \in \mathbb{R}^N$ such that $\phi_i(N, v)$ is interpreted as the payoff allocated to player $i \in N$.

3. FREIGHT TRANSPORTATION GAMES

Let $\mathcal{N}^1 = \{N \in \mathcal{N} : 1 \in N\}$. In our context, $N \in \mathcal{N}^1$ is the set of players with 1 represents a mining firm and

$$N' = N \setminus \{1\}$$

representing the set of local communities with a fixed order.

Consider the network where all players are in a line. We assume, w.l.o.g., the order of the players in the line is given by its cardinally, i.e. $2, 3, \dots, n$ when $N = \{1, 2, \dots, n\}$. A coalition

$S \subseteq N'$ of communities is *connected* if the sub-network restricted to agents in S has a single component. We denote by $\mathcal{C}$ the set of connected coalitions. Formally,

$$\mathcal{C} = \{S \subseteq N' : i, k \in S,\ i < j < k \Rightarrow j \in S\}.$$

Notice that, by definition, $\emptyset, N', \{i\} \in \mathcal{C}$ for all $i \in N'$.

Since the communities are located on a linear route, we can work with the notion of consecutive communities, formally defined as follows:

Definition 1. *Coalitions $S, T \in \mathcal{C}$ are* consecutive *if $S \cup T \in \mathcal{C}$ and $S \cap T = \emptyset$.*

Let $\mathcal{A} \subseteq \mathcal{C}$ denote the set of consecutive communities that have an alternative route option. For notational convenience, we assume $\emptyset \in \mathcal{A}$. An example will clarify this notion:

Example 1. *Assume $N' = \{2, 3, 4, 5\}$. A set of consecutive communities is given by $\mathcal{A} = \{\{2, 3\}, \{3\}, \{4\}, \{5\}, \{4, 5\}\}$. This means that there exists an alternative road that can be built circumventing communities 2 and 3 altogether, and another alternative circumventing communities 4 and 5 altogether. Moreover, there is also another alternative circumventing community 3, another one circumventing community 4, and another one circumventing community 5, so that it is possible to avoid, say, only community 4. As opposed, there is no alternative to avoid community 2 alone. If community 2 does not cooperate, then cooperation from community 3 alone is not enough.*

Definition 2. *A* path compatible with $\mathcal{A}$ *is a pair $(\mathcal{P}, f)$ where*

- *$\mathcal{P} = \left\{P^1, \ldots, P^k\right\} \subset \mathcal{C}$ such that $P^l \in \mathcal{A}$ whenever $\left|P^l\right| > 1$, P^{l-1} and P^l are consecutive for all $l = 2, \ldots, k$, and $\bigcup_{l=1}^{k} P^l = N'$.*
- *$f : \mathcal{P} \to \{col, alt\}$ is a function that assigns either col (collaborate) or alt (alternative) to each $P^l \in \mathcal{P}$ and such that $f\left(P^l\right) = alt$ whenever $\left|P^l\right| > 1$ and $f\left(P^l\right) = col$ whenever $P^l \notin \mathcal{A}$.*

Let $\mathbb{P}(\mathcal{A})$ denote the set of paths compatible with $\mathcal{A}$. Notice that $\mathbb{P}(\mathcal{A})$ is always nonempty, as $(\mathcal{P}, f)$ with $\mathcal{P} = \{\{2\}, \{3\}, \ldots, \{n\}\}$ and $f(\{i\}) = col$ for all $i \in N'$ always belongs to it. Some compatible paths in the previous example are the following:

Example 2. *Assume $N' = \{2, 3, 4, 5\}$. A set of consecutive communities is*

$$\mathcal{A} = \{\{2, 3\}, \{3\}, \{4\}, \{5\}, \{4, 5\}\}.$$

The following are paths compatible with $\mathcal{A}$:

- *$(\mathcal{P}^1, f^1)$ with $\mathcal{P}^1 = \{\{2\}, \{3\}, \{4, 5\}\}$, $f^1(\{2\}) = col$, $f^1(\{3\}) = alt$, $f^1(\{4, 5\}) = alt$. In this path, the mining freight goes through community 2, and avoids communities 3, 4, and 5.*
- *$(\mathcal{P}^2, f^2)$ with $\mathcal{P}^2 = \{\{2\}, \{3\}, \{4, 5\}\}$, $f^2(\{2\}) = col$, $f^2(\{3\}) = col$, $f^2(\{4, 5\}) = alt$. In this path, the mining freight goes through communities 2 and 3, and avoids communities 4 and 5.*
- *$(\mathcal{P}^3, f^3)$ with $\mathcal{P}^3 = \{\{2, 3\}, \{4\}, \{5\}\}$, $f^3(\{2, 3\}) = alt$, $f^3(\{4\}) = col$, $f^3(\{5\}) = alt$. In this path, the mining freight goes through community 4, and avoids communities 2, 3, and 5.*

Definition 3. *A freight transport problem is a tupla $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ where $N \in \mathcal{N}^1$, $E > 0$, $c \in \mathbb{R}_+^N$, $\mathcal{A} \subset \mathcal{C}$, and $a \in \mathbb{R}_+^{\mathcal{A}}$ ($a_\emptyset = 0$) satisfies*

$$a_A + c(B \setminus A) \leq a_B \tag{1}$$

for all $A, B \in \mathcal{A}$ with $A \subset B$, and

$$a_A \leq a_{A_1} + a_{A_2} + \cdots + a_{A_k} \tag{2}$$

whenever $A, A_1, A_2, \ldots, A_k \in \mathcal{A}$ and $A = \bigcup_{l=1}^{k} A_l$.

In particular, a is a vector whose coordinates are the costs of using alternative routes avoiding each coalition in $\mathcal{A}$. The reason for (1) and (2) is that communities in $A \in \mathcal{A}$ would be irrelevant if we allow a_A to be too high so that they could then be avoided at no additional cost by the mining firm, either by using a supra-alternative (1) or several intra-alternatives (2).

We analyze in depth this and the rest of the components of a freight transportation problem:

1. N is the set of players with $1 \in N$ the mining firm and $N' = N \setminus \{1\}$ the ordered communities.

2. E represents the total benefit generated by freight transportation, i.e., the benefit the mining industry gets from the ore once arrived at its destination.

3. The vector cost $c \in \mathbb{R}_+^{N'}$ represents the costs that the transportation of minerals affects the respective local community. Hence, a community $i \in N'$ suffers a cost c_i only if the mineral goes through its land.

4. $\mathcal{A}$ represents the set of communities that have an alternative route avoiding the respective community.

5. Finally, $a \in \mathbb{R}_+^{\mathcal{A}}$, such that a_A with $A \in \mathcal{A}$, is the cost of building/using the respective alternative route that avoids going thought players in A.

Example 3. *Let $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ with $N = \{1, 2, 3, 4, 5\}$, $E = 11$, $c_i = 0$ for all $i \in N'$, $\mathcal{A} = \{\{2, 3\}, \{3\}, \{4\}, \{5\}, \{4, 5\}\}$, $a_{\{2,3\}} = 4$, $a_{\{3\}} = a_{\{4\}} = 2$, $a_{\{5\}} = 4$ and $a_{\{4,5\}} = 5$. See Figure 1. It is straightforward to check that (1) and (2) are satisfied.*

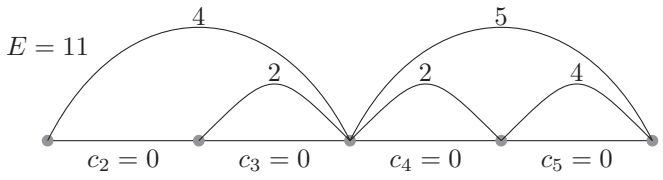


Figure 1: Example of a freight transportation problem.

Example 4. *Let $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ with $N = \{1, 2, 3, 4, 5\}$, $E = 8$, $c_2 = c_3 = c_4 = 1$, $c_5 = 0$, $\mathcal{A} = \{\{2\}, \{3\}, \{4\}, \{4, 5\}, \{3, 4, 5\}\}$, $a_{\{2\}} = a_{\{3\}} = 2$, $a_{\{4\}} = 4$, $a_{\{4,5\}} = 4$ and $a_{\{3,4,5\}} = 6$. See Figure 2. It is straightforward to check that (1) and (2) are satisfied.*

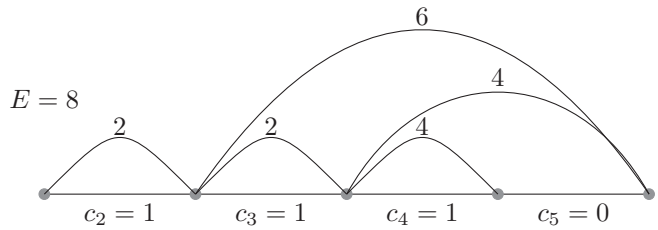


Figure 2: Example of a freight transportation problem.

We also assume the next conditions:

Assumption 1 We assume that there is enough benefit of cooperation even without collaboration from the local communities, i.e., there exists a partition of N', $\mathcal{A}' = \{A_N^1, \ldots, A_N^k\} \subseteq \mathcal{A}$ such that $E \geq \sum_{A \in \mathcal{A}'} a_A$.

In other words, there exists at least one alternative that avoids all communities. Let denote as $\mathcal{A}^{(a)} \subset \mathcal{A}$ a partition with minimum $\sum_{A \in \mathcal{A}^{(a)}} a_A$.

Assumption 2. The alternative options can not properly overlap, i.e., given $i < j \leq k < l \in N'$ and $\{i, i+1, ..., k\} \in \mathcal{A}$, then $\{j, j+1, ..., l\} \notin \mathcal{A}$.

In other words, if there exists an alternative between i and k, then, there is no alternative option between j and l.

Notice that both freight transportation problems given in Example 3 and Example 4, respectively, satisfy Assumption 1 and Assumption 2.

Proposition 1. *Assumption 2 holds if and only if $S \cap T \in \{S, T, \emptyset\}$ for all $S, T \in \mathcal{A}$.*

Proof available on request.

Proposition 2. *Under Assumption 1 and Assumption 2, there exists a unique coarsest partition $\mathcal{A}'$ of N' formed by elements in $\mathcal{A}$.*

Proof available on request.

Corollary 1. *For each $\mathcal{F} = (N, E, c, \mathcal{A}, a)$,*

$$E' = \sum_{A \in \mathcal{A}^{(a)}} a_A = \sum_{A \in \mathcal{A}'} a_A.$$

Proof available on request.

Under Corollary 1, we can assume $A^{(a)} = A'$ independent of a.

Notice that Assumption 2 is satisfied in the alternative sets given in Example 1 and Example 2.

Let $\mathfrak{F}$ be the class of all freight transportation problems satisfying Assumption A1 and Assumption 2.

Given $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$, we associate a TU-game $\left(N, v^{\mathcal{F}}\right)$, where N is the set of players including the mining firm and local communities and $v^{\mathcal{F}}$ is a characteristic function defined as follows. Given $S \subseteq N$ with $1 \notin S$, $v^{\mathcal{F}}(S) = 0$. Assume now $1 \in S$. Then,

$$v^{\mathcal{F}}(S) = \max_{(\mathcal{P}, f) \in \mathbb{P}^S(\mathcal{A})} \left\{ E - \sum_{P^l \in \mathcal{P}: f(P^l) = alt} a_{P^l} - \sum_{P^l \in \mathcal{P}: f(P^l) = col} c\left(P^l\right) \right\}$$

where

$$\mathbb{P}^S(\mathcal{A}) = \left\{ (\mathcal{P}, f) \in \mathbb{P}(\mathcal{A}) : P^l \in \mathcal{P}, P^l \not\subseteq S \Rightarrow f(P^l) = alt \right\}.$$

Clearly, $S \subset T$ implies $v^{\mathcal{F}}(S) \leq v^{\mathcal{F}}(T)$, i.e., $v^{\mathcal{F}}$ is a monotonic TU-game.

Example 5. *The following Tables 1 and 2 show the worth of some coalitions in the cooperative games associated with the freight transportation problems given in Example 3 and Example 4, respectively:*

Table 1: The worth of some coalitions in Example 3.

S	$v^{\mathcal{F}}(S)$	Optimal $\mathcal{P}$	Optimal f
$\{1\}$	2	$\{\{2,3\},\{4,5\}\}$	$f(\{2,3\}) = f(\{4,5\}) = alt$
$\{1,2\}$	4	$\{\{2\},\{3\},\{4,5\}\}$	$f(\{2\}) = col, f(\{3\}) = f(\{4,5\}) = alt$
$\{1,2,3\}$	6	$\{\{2\},\{3\},\{4,5\}\}$	$f(\{2\}) = f(\{3\}) = col, f(\{4,5\}) = alt$
$\{1,2,4,5\}$	9	$\{\{2\},\{3\},\{4\},\{5\}\}$	$f(\{2\}) = f(\{4\}) = f(\{5\}) = col, f(\{3\}) = alt$
N	11	$\{\{2\},\{3\},\{4\},\{5\}\}$	$f(\{i\}) = col \ \ \forall i \in N'$

Table 2: The worth of some coalitions in Example 4.

S	$v^{\mathcal{F}}(S)$	Optimal $\mathcal{P}$	Optimal f
$\{1\}$	0	$\{\{2\},\{3,4,5\}\}$	$f(\{2\}) = f(\{3,4,5\}) = alt$
$\{1,2\}$	1	$\{\{2\},\{3,4,5\}\}$	$f(\{2\}) = col,\ f(\{3,4,5\}) = alt$
$\{1,2,3\}$	2	$\{\{2\},\{3\},\{4,5\}\}$	$f(\{2\}) = f(\{3\}) = col, f(\{4,5\}) = alt$
$\{1,2,4,5\}$	4	$\{\{2\},\{3\},\{4\},\{5\}\}$	$f(\{2\}) = f(\{4\}) = f(\{5\}) = col, f(\{3\}) = alt$
N	5	$\{\{2\},\{3\},\{4\},\{5\}\}$	$f(\{i\}) = col \ \ \forall i \in N'$

We also define the core of $\mathcal{F} \in \mathfrak{F}$ as follows:

$$Core(\mathcal{F}) = Core\left(N, v^{\mathcal{F}}\right).$$

Our aim is to study whether the core of a freight transportation problem is nonempty, and, if so, how to find a core allocation.

4. ESSENTIAL PROBLEMS

To study the core in freight transportation games, we use the concept of *essential freight transportation problems*, defined as follows:

Definition 4. *Given $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$, we say that $\mathcal{F}$ is* essential *if*

1. $c_i = 0$ *for all* $i \in N'$,
2. $a_A < a_B$ *for all* $A, B \in \mathcal{A}$ *with* $A \subset B$, *and*
3. $a_{A_0} < \sum_{l=1}^{k} a_{A_l}$ *for all* $A_0, A_1, \ldots, A_k \in \mathcal{A}$ *with* $A_0 = \bigcup_{l=1}^{k} A_l$.

It is not difficult to check that the freight transportation problem defined in Example 3 is essential. By contrast, the freight transportation problem presented in Example 4 is not essential, because it does not satisfy any of the three conditions: $c_i \neq 0$ for all $i \in N'$, $a_{\{4\}} = 4 = 4 = a_{\{4,5\}}$, and $a_{\{3,4,5\}} = 6 = a_{\{3\}} + a_{\{4,5\}}$. However, there exists an essential problem $\mathcal{F}^*$ that generates the same cooperative game as the problem defined in Example 4. It is given by $\mathcal{F}^* = (N, E^*, c^*, \mathcal{A}^*, a^*)$ with $N = \{1, 2, 3, 4, 5\}$, $E^* = 5$, $c_i^* = 0$ for all $i \in N'$, $\mathcal{A}^* = \{\{2\}, \{3\}, \{4, 5\}\}$, $a^*_{\{2\}} = a^*_{\{3\}} = 1$, and $a^*_{\{4,5\}} = 3$ (see Figure 3).

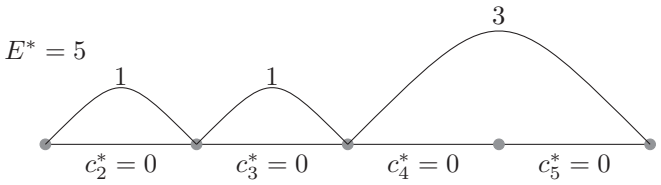


Figure 3: Essential freight transportation problem with the same characteristic function as those given in Example 4.

This result is general, as shown in Proposition 4. Another advantage of essential problems is that the optimal paths are easily determined, as the next result shows.

Proposition 3. *If $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$ is essential, then $v^{\mathcal{F}}(N) = E$ and $a_A = E - v^{\mathcal{F}}(N \setminus A)$ for all $A \in \mathcal{A}$. In particular, for any $A \in \mathcal{A}$, the (unique) optimal path for $v^{\mathcal{F}}(N \setminus A)$ is $(\mathcal{P}, f) \in \mathbb{P}^{N \setminus A}(\mathcal{A})$ given by $\mathcal{P} = \{A\} \cup \{\{i\}\}_{i \in N' \setminus A}$, $f(A) = alt$, and $f(\{i\}) = col$ for all $i \in N' \setminus A$.*

Proof available on request.

Proposition 4. *For all $\mathcal{F} \in \mathfrak{F}$, there exists a unique $\mathcal{F}^* \in \mathfrak{F}$ essential such that $v^{\mathcal{F}} = v^{\mathcal{F}^*}$.*

Proof available on request.

5. THE CORE

In this section, we study the core of freight transportation problems. In particular, we study whether it is always possible to find a core allocation, i.e., an agreement which no coalition of players has incentives to reject.

The next results show that the answer is affirmative. The core is always nonempty in a freight transportation problem.

Theorem 1. *Given $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$, the core associated to $\mathcal{F}$ is:*

$$Core(\mathcal{F}) = \left\{x \in \mathbb{R}_+^N : x(N) = E - c(N'),\ x(A) \leq a_A - c(A)\ \forall A \in \mathcal{A}\right\}.$$

Proof available on request.

Corollary 2. *The core of any freight transport problem satisfying Assumption 1 and Assumption 2 is always nonempty.*

Proof available on request.

6. A FAMILY OF CORE ALLOCATION RULES

In this section, we propose a family of core allocation rules. As the first step, we define the concept of levels structure, a concept first formalized in the context of cooperative game theory by Winter (1989).

Definition 5. *We define a levels structure over N' as a finite sequence $\mathfrak{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m)$ of partitions of N' such that, for each $A \in \mathcal{A}_l$, $l > 1$, there exists $\mathcal{B} \subseteq \mathcal{A}_{l-1}$ with $A = \bigcup_{B \in \mathcal{B}} B$.*

Lemma 1. *If $\mathcal{F} \in \mathfrak{F}$ is essential, then there exists a unique levels structure $\mathfrak{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_m)$ over N' such that:*

1. *All partitions A_l are different.*
2. *$A \in \mathcal{A} \cup \{N'\} \cup \{\{i\}\}_{i \in N'}$ if and only if there exists $l = 1, \ldots, m$ such that $A \in \mathcal{A}_l$.*
3. *If $A \in \mathcal{A}_l$, $l > 1$, and $\mathcal{B} \subseteq \mathcal{A}_{l-1}$ is such that $A = \bigcup_{B \in \mathcal{B}} B$, then either $|\mathcal{B}| > 1$ or $|A| = 1$.*

In particular, $\mathcal{A}_1 = \{\{i\}\}_{i \in N'}$ and $\mathcal{A}_m = \{N'\}$.

Proof. The proof is constructive. The first partition is defined as $\mathcal{A}_m = \{N'\}$. Assume we have defined $\mathcal{A}_{l+1}, \ldots, \mathcal{A}_m$. We define

$$\widetilde{\mathcal{A}}_l = \{A \in \mathcal{A} : A \subset B \in \mathcal{A} \implies B \in \mathcal{A}_{l'} \text{ for some } l' > l\}$$

and

$$\mathcal{A}_l = \widetilde{\mathcal{A}}_l \cup \{\{i\}\}_{i \in N' \setminus \bigcup_{A \in \widetilde{\mathcal{A}}_l} A}.$$

The process continues until all $A \in \mathcal{A}$ belong to some $\mathcal{A}_l$, so that condition 1 is satisfied. In this case, $\widetilde{\mathcal{A}}_l = \widetilde{\mathcal{A}}_1 = \emptyset$, so that $\mathcal{A}_1 = \{\{i\}\}_{i \in N'}$. □

Example 6. *Assume $\mathcal{A} = \{\{2\}, \{3\}, \{4\}, \{2,3\}, \{2,3,4\}\}$. Then, $\mathfrak{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ where $\mathcal{A}_1 = \{\{2\}, \{3\}, \{4\}\}$, $\mathcal{A}_2 = \{\{2,3\}, \{4\}\}$ and $\mathcal{A}_3 = \{N'\}$.*

A *rule* ϕ is a function that assigns to each problem $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$ a payoff allocation $\phi(\mathcal{F}) \in \mathbb{R}^N$.

We define a family of rules as follows: Let $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ and let $\mathcal{F}^* = (N, E^*, c^*, \mathcal{A}^*, a^*)$ be the (unique) essential problem associated with $\mathcal{F}$ as given by Proposition 4. Let $\mathfrak{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m)$ be the unique level structure associated to $\mathcal{F}^*$ as given in Lemma 1. Given $i \in N'$ and $l \in \{1, \ldots, m\}$, let $A_l^i \in \mathcal{A}_l^*$ be the (unique) coalition in level l that contains agent i. Obviously, $i \in A_1^i \subseteq A_2^i \subseteq \cdots \subseteq A_m^i$. Let $\mathcal{A}' \subseteq \mathcal{A}$ be the partition of N' so that $v^{\mathcal{F}}(\{1\}) = E - \sum_{A \in \mathcal{A}^{(a)}} a_A$ (as given by Assumption A1). Given $\theta \in \mathbb{R}$, we define the rule ϕ^θ as follows:

$$\phi_1^\theta(\mathcal{F}) = E - \theta E' - (1 - \theta)\, c(N')$$

and

$$\phi_i^\theta(\mathcal{F}) = \theta\,(E' - c(N')) \prod_{l=1}^{m-1} \frac{\widetilde{a}_{A_l^i}}{\sum_{A \in \mathcal{A}_l : A \subseteq \mathcal{A}_{l+1}^i} (\widetilde{a}_A)} \tag{3}$$

for each $i \in N'$, where A_l^i is the (only) coalition in $\mathcal{A}_l$ that contains player i, and

$$\widetilde{a}_A = \min_{B \in \mathcal{A}^* : A \subseteq B} a_B^*.$$

Notice that $\widetilde{a}_A = a_A$ whenever $a \in \mathcal{A}^*$, and otherwise $A = \{i\}$ for some $i \in N'$, so that $\widetilde{a}_{\{i\}} = \min_{B \in \mathcal{A}^* : i \in B} a_B^*$. Moreover, for each $i \in N'$, there exists some $l^i \in \{1, \ldots, m\}$ such that $A_{l^i}^i \in \mathcal{A}^*$ and $A_l^i = \{i\}$ for all $l < l^i$, so that $\widetilde{a}_{\{i\}} = a^*_{A_{l^i}^i}$. Under Proposition 3, we deduce that

$$\widetilde{a}_{\{i\}} = v^{\mathcal{F}}(N) - v^{\mathcal{F}}\left(N \setminus A_{l^i}^i\right).$$

We now define some reasonable properties, as follows:

Core selection (CS) Given $\mathcal{F}$, $\phi(\mathcal{F}) \in Core(\mathcal{F})$.

Core selection is a property that concerns the stability of a solution.

For the next property, we consider the problem that arises when a group of consecutive communities merge. This property prevents communities from manipulating their outcome by merging or splitting. It is a very relevant property in situations where the identity of the communities is unclear, as they can, for example, associate at different local levels or administrative levels (village, town hall, region, etc.).

Definition 6. *Two coalitions A, $A' \in \mathcal{A}$ are mergeable if they belong to the same supra-coalitions, i.e., for all $B \in \mathcal{A}$, $A \subsetneq B$ if and only if $A' \subsetneq B$.*

Independence of Merging of Mergeable Alternatives (IMMA) Given k pairwise disjoint consecutive and mergeable coalitions $A_1, \ldots, A_k \in \mathcal{A}$ with $i \in A = \bigcup_{l=1}^{k} A_l$, then, $\phi_j(\mathcal{F}) = \phi_j(\mathcal{F}^{A,i})$ for all $j \in N' \setminus A$, where

$$\mathcal{F}^{A,i} = \left((N \setminus A) \cup \{i\}, E, c^{A,i}, \mathcal{A}^{A,i}, a^{A,i}\right)$$

is given by

- $c_i^{A,i} = \sum_{j \in A} c_j$, $c_j^{A,i} = c_j$ otherwise,
- $\mathcal{A}^{A,i} = \{B \in \mathcal{A} : A \cap B = \emptyset\} \cup \{(B \setminus A) \cup \{i\} : B \in \mathcal{A}, A \subseteq B\} \cup \{\{i\}\}$,
- $a_B^{A,i} = a_B$ if $A \cap B = \emptyset$, $a_{(B \setminus A) \cup \{i\}}^{A,i} = a_B$ if $A \subsetneq B$, and

$$a_{\{i\}}^{A,i} = \begin{cases} a_A & \text{if } A \in \mathcal{A} \\ \sum_{l=1}^{k} a_{A_l} & \text{if } A \notin \mathcal{A}. \end{cases}$$

IMMA states that, if a group of communities with an essential alternative merge, no other agent should be affected. Alternatively, if a community with an essential alternative splits into new consecutive communities, having the new group an essential alternative, no other agent should be affected.

Equivalence (EQUI). Given $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$, $\phi(\mathcal{F}) = \phi\left(\widetilde{\mathcal{F}}\right)$, where $\widetilde{\mathcal{F}} = \left(N, E, c, \widetilde{\mathcal{A}}, \widetilde{a}\right)$ is defined as $\widetilde{\mathcal{A}} = \mathcal{A} \cup \{\{i\}\}_{i \in N'}$, $\widetilde{a}_A = a_A$ for all $A \in \mathcal{A}$, and

$$\widetilde{a}_{\{i\}} \geq \min_{A \in \mathcal{A} : i \in A} \{a_A - c(A \setminus \{i\})\}$$

for all $\{i\} \notin \mathcal{A}$.

EQUI states that the payoffs do not change if we assume that communities with no feasible alternative (i.e. those $i \in N'$ such that $\{i\} \notin \mathcal{A}$) do have an alternative, but it is so expensive that it is not worthy to use it.

Two communities $i, j \in N'$ are *symmetric* if they have the same cost, they belong to the same supra-coalitions, and their alternative costs (if any) are equal, i.e.,

- $c_i = c_j$,
- if $\{i\} \subsetneq A \in \mathcal{A}$, then $j \in A$; analogously, if $\{j\} \subsetneq A \in \mathcal{A}$, then $i \in A$, and
- if $\{i\} \in \mathcal{A}$, then $\{j\} \in \mathcal{A}$ and $a_{\{i\}} = a_{\{j\}}$; analogously, if $\{j\} \in \mathcal{A}$, then $\{i\} \in \mathcal{A}$ and $a_{\{i\}} = a_{\{j\}}$.

Symmetry (SYM): If $i, j \in N'$ are symmetric, then $\phi_i(\mathcal{F}) = \phi_j(\mathcal{F})$.

Additivity (ADD). Given $\mathcal{F}^1 = \left(N, E^1, c^1, \mathcal{A}, a^1\right), \mathcal{F}^2 = \left(N, E^2, c^2, \mathcal{A}, a^2\right) \in \mathfrak{F}$, $\phi(\mathcal{F}^1 + \mathcal{F}^2) = \phi(\mathcal{F}^1) + \phi(\mathcal{F}^2)$, where

$$\mathcal{F}^1 + \mathcal{F}^2 = \left(N, E^1 + E^2, c^1 + c^2, \mathcal{A}, a^1 + a^2\right).$$

In order to illustrate ADD, assume that a mining firm transports copper and gold. We consider that E is the benefit generated by copper transportation and E' by gold transportation in the mining transportation context, c and c' represent costs from different sources like air and land pollution, analogously for a and a'. Then, when computing the compensations that the mining firm should provide to the communities, ADD states that there is no difference between considering both problems separately or in aggregate.

A special class of problems are those where the cost of any alternative coincides with the cost of the communities that such an alternative avoids. Namely, $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ is such that $a_A = c(A)$ for all $A \in \mathcal{A}$. Next property forces additivity only when one of the problems satisfies such condition:

Community Cost Separation (CCS) Given $\mathcal{F}^1 = (N, E^1, c^1, \mathcal{A}, a^1)$, $\mathcal{F}^2 = (N, E^2, c^2, \mathcal{A}, a^2) \in \mathfrak{F}$ such that $a^2_A = c^2(A)$ for all $A \in \mathcal{A}$, then

$$\phi\left(\mathcal{F}^1 + \mathcal{F}^2\right) = \phi(\mathcal{F}^1) + \phi(\mathcal{F}^2).$$

Lemma 2. *Let ϕ be a rule that satisfies CS, IMMA and SYM. Then, for each $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ with $c_i = 0$ and $\{i\} \in \mathcal{A}$ for all $i \in N'$:*

$$\phi_i(\mathcal{F}) = (E - \phi_1(\mathcal{F})) \prod_{l=1}^{m-1} \frac{a_{A^i_l}}{\sum_{A \in \mathcal{A}_l : A \subset \mathcal{A}^i_{l+1}} (a_A)} \tag{4}$$

for all $i \in N'$, where $\mathfrak{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m)$ is the unique level structure associated to $\mathcal{F}^*$ as given in Lemma 1 and $A^i_l \in \mathcal{A}_l$ is the (unique) coalition in level l that contains agent i.

Proof of Lemma 2 is available on request. A sketch of the proof is in the following example.

Example 7. *Let $\mathcal{F} = (N, E, c, \mathcal{A}, a)$ with $N = \{1, 2, 3, 4\}$, $E = 10$, $c_i = 0$ for all $i \in N'$, $\mathcal{A} = \{\{2\}, \{3\}, \{4\}, \{2, 3\}, \{2, 3, 4\}\}$, $a_{\{2\}} = 3$, $a_{\{3\}} = 4$, $a_{\{4\}} = 3$, $a_{\{2,3\}} = 5$, and $a_{\{2,3,4\}} = 6$. See Figure 4.*

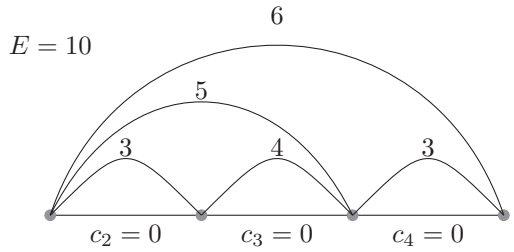


Figure 4: Example of a freight transportation problem, where communities 2 and 3 are mergeable, and so are coalitions $\{2, 3\}$ and $\{4\}$.

In this example, $m = 3$ and $\mathfrak{A} = (\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3)$ where

$$\mathcal{A}_1 = \{\{2\}, \{3\}, \{4\}\}$$
$$\mathcal{A}_2 = \{\{2, 3\}, \{4\}\}$$
$$\mathcal{A}_3 = \{\{2, 3, 4\}\}.$$

Assume ϕ satisfies CS, IMMA and SYM. We need to prove (4), which in this example means $\phi_2(\mathcal{F}) = (E - \phi_1(\mathcal{F}))\frac{3}{7}\frac{5}{8}$, $\phi_3(\mathcal{F}) = (E - \phi_1(\mathcal{F}))\frac{4}{7}\frac{5}{8}$, and $\phi_4(\mathcal{F}) = (E - \phi_1(\mathcal{F}))\frac{3}{3}\frac{3}{8}$.

Under CS, $\phi_1(\mathcal{F})+\phi_2(\mathcal{F})+\phi_3(\mathcal{F})=E-\phi_1(\mathcal{F})$. Communities 2 and 3 are mergeable. Hence, under IMMA, we can merge them and work with the reduced problem $\mathcal{F}'=\mathcal{F}^{\{2,3\},2}$ with communities $\{2,4\}$ and $a'_{\{2\}}=5$, $a'_{\{4\}}=3$, and $a'_{\{2,4\}}=6$. We apply again IMMA to split these new communities and work with the extended problem with eight symmetric communities, each of them with $a''_{\{i\}}=1$. Under SYM, each split community receives $\frac{E-\phi_1(\mathcal{F})}{8}$. Under IMMA, we can remerge the eight communities into 2 and 4, so that $\phi_2(\mathcal{F}')=5\frac{E-\phi_1(\mathcal{F})}{8}$ and $\phi_4(\mathcal{F})=\phi_4(\mathcal{F}')=3\frac{E-\phi_1(\mathcal{F})}{8}$. We repeat the same reasoning with IMMA and SYM applied to coalition $\{2\}$ alone, so that we conclude that $\phi_2(\mathcal{F})=5\frac{E-\phi_1(\mathcal{F})}{8}\frac{3}{7}$ and $\phi_3(\mathcal{F})=5\frac{E-\phi_1(\mathcal{F})}{8}\frac{4}{7}$, as desired.

Theorem 2. *A rule ϕ satisfies CS, IMMA, EQUI, SYM, and CCS if and only if there exists $\theta\in[0,1]$ such that $\phi=\phi^\theta$.*

Proof available on request.

These properties are independent. We define rules that satisfy all the properties but exactly one. Given $\mathcal{F}=(N,E,c,\mathcal{A},a)\in\mathfrak{F}$:

- Rule ϕ^θ with $\theta\notin[0,1]$ satisfies all the properties but *CS*.
- Let $\overline{\phi}$ defined as
$$\overline{\phi}(\mathcal{F})=\begin{cases}\phi^0(\mathcal{F}) \text{ if } |N'|>1\\ \phi^1(\mathcal{F}) \text{ if } |N'|=1.\end{cases}$$
This rule satisfies all the properties but IMMA.
- Let $\overline{\phi}^1$ defined as
$$\overline{\phi}^1_1(\mathcal{F})=E-E'$$
and, for each $i\in N'$,
$$\overline{\phi}^1_i(\mathcal{F})=(E-c(N'))\prod_{l=1}^{m-1}\frac{\overline{a}_{A^i_l}}{\sum_{A\in\mathcal{A}_l:A\subset\mathcal{A}^i_{l+1}}(\overline{a}_A)}$$
where $\overline{a}_A=a_A$ for all $A\in\mathcal{A}^*$ and $\overline{a}_i=0$ for all $\{i\}\in\mathcal{A}^*\setminus\mathcal{A}$. This rule satisfies all the properties but EQUI.
- Let $\phi^{(1st)}$ defined as the rule that gives, among all core allocations of the associated essential problem, the one that assigns the maximum possible to the first community, then the maximum possible to the second one, and so on, up to E'. This rule satisfies all the properties but SYM.
- The rule defined as $\phi(\mathcal{F})=\phi^{\frac{E'}{E}}(\mathcal{F})$ satisfies all the properties but CCS.

Moreover, ϕ^0 also satisfies ADD:

Proposition 5. *ϕ^0 satisfies ADD.*

Proof. Let $\mathcal{F}^1=(N,E^1,c^1,\mathcal{A},a^1)$ and $\mathcal{F}^2=(N,E^2,c^2,\mathcal{A},a^2)$ given as in the definition of ADD. Then,
$$\begin{aligned}\phi^0_1(\mathcal{F}^1+\mathcal{F}^2)&=(E^1+E^2)-\left(c^1(N')+c^2(N')\right)\\&=E^1-c^1(N')+E^2-c^2(N')\\&=\phi^0_1(\mathcal{F}^1)+\phi^0_1(\mathcal{F}^2)\end{aligned}$$
and $\phi^0_i(\mathcal{F}^1+\mathcal{F}^2)=0=\phi^0_2(\mathcal{F}^1)+\phi^0_2(\mathcal{F}^2)$ for each $i\in N'$. □

For $\theta>0$, ϕ^θ does not satisfy ADD, as the next example shows.

Example 8. *Let* $\mathcal{F} = (N, E, c, \mathcal{A}, a) \in \mathfrak{F}$ *given by* $N = \{1, 2, 3\}$, $E = 4$, $c_2 = c_3 = 0$, $\mathcal{A} = \{\{2\}, \{3\}, \{2, 3\}\}$, $a_{\{2\}} = 2$, $a_{\{3\}} = 3$, *and* $a_{\{2,3\}} = 4$. *Let* $\mathcal{F}^1 = (N, E^1, c, \mathcal{A}, a^1) \in \mathfrak{F}$ *given by* $E^1 = 2$, $a^1_{\{2\}} = 1$, *and* $a^1_{\{3\}} = a^1_{\{2,3\}} = 2$. *Let* $\mathcal{F}^2 = (N, E^2, c, \mathcal{A}, a^2) \in \mathfrak{F}$ *given by* $E^2 = 2$, $a^2_{\{2\}} = a^2_{\{3\}} = 1$, *and* $a^2_{\{2,3\}} = 2$. *Clearly,* $\mathcal{F} = \mathcal{F}^1 + \mathcal{F}^2$. *Moreover,*

$$\phi^\theta(\mathcal{F}) = \left(4 - 4\theta, \frac{8}{5}\theta, \frac{12}{5}\theta\right).$$

On the other hand,

$$\phi^\theta(\mathcal{F}^1) = \left(2 - 2\theta, \frac{2}{3}\theta, \frac{4}{3}\theta\right)$$

and

$$\phi^\theta(\mathcal{F}^2) = (2 - 2\theta, \theta, \theta).$$

Hence, $\phi^\theta(\mathcal{F}) = \phi^\theta(\mathcal{F}^1 + \mathcal{F}^1) \neq \phi^\theta(\mathcal{F}^1) + \phi^\theta(\mathcal{F}^2)$ *whenever* $\theta > 0$.

Theorem 3. *A rule* ϕ *satisfies CS, IMMA, SYM, and ADD if and only if* $\phi = \phi^0$.

Proof available on request.

7. CONCLUSIONS

In this paper, we assess the potential of game theory through cooperative games applied to the freight transportation of minerals. We assess that benefits/costs would result from considering the bargaining power of communities to avoid the use of their land. It is a methodological contribution that analyses road use management for freight transportation. The result is positive in the sense that we show that cooperative games are indeed a useful tool. We consider it important to apply this model to different study regions, such as the Apurimac-Cusco area in Peru. It is advisable for a practical application where we use cooperative game theory to state possible agreements between the mining firms and local communities.

This study shows a framework for allocating compensations to communities based on cooperative game theory, taking into account the principle of stability. We show that it is possible to establish compensation rules that assure stability for local communities. Specifically, we define several reasonable properties in the context of freight transport problems and propose a parametric family of solutions, which satisfy each property. One of the properties is core selection, which allows us to assign stable solutions.

ACKNOWLEDGMENTS

Juan Vidal-Puga acknowledges financial support from Grant TED2021-130241A-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. The authors also acknowledge financial support from FONDECYT (Now PROCIENCIA) through "Proyecto de Investigación Básica Convenio 061-2021 FONDECYT". Financial support from the Spanish Ministerio de Ciencia e Innovación through grant PID2020-113440GB-I00 and Xunta de Galicia (Consellería de Cultura, Educación, Formación Profesional e Universidades) through grant GPC-ED431B 2022/03 is also gratefully acknowledged.

REFERENCES

Alcalde-Unzu, J., Gómez-Rúa, M., and Molis, E. (2021) Allocating the costs of cleaning a river: expected responsibility versus median responsibility. International Journal of Game Theory, 50, 185–214.

Gillies, D. (1959). Solutions to general non-zero sum games. In Tucker, A. and Luce, R., editors, Contributions to the theory of games, volume IV of Annals of Mathematics Studies, chapter 3, pages 47–85. Princeton UP, Princeton.

Gustafsson, M.-T. and Scurrah, M. (2019) Strengthening subnational institutions for sustainable development in resource-rich states: Decentralized land-use planning in Peru. World Development, 119, 133–144.

Lin, B., Zhao, Y., Lin, R., and Liu, C. (2021). Integrating traffic routing optimization and train formation plan using simulated annealing algorithm. Applied Mathematical Modelling, 93, 811–830.

Liu, C., Du, Y., Wong, S., Chang, G., and Jiang, S. (2020). Eco-based pavement lifecycle maintenance scheduling optimization for equilibrated networks. Transportation Research Part D: Transport and Environment, 86, 102471.

Ma, J., Li, D., Cheng, L., Lou, X., Sun, C., and Tang, W. (2018). Link restriction: Methods of testing and avoiding Braess paradox in networks considering traffic demands. Journal of Transportation Engineering, Part A: Systems, 144, 04017076.

Sun, P., Hou, D., and Sun, H. (2019). Responsibility and sharing the cost of cleaning a polluted river. Mathematical Methods of Operations Research, 89, 143–156.

van den Brink, R., He, S., and Huang, J.-P. (2018). Polluted river problems and games with a permission structure. Games and Economic Behavior, 108, 182–205.

Winter, E. (1989) A value for cooperative games with levels structure of cooperation. International Journal of Game Theory 18(2), 227-240.

Ye, Y., Wang, H., Zhang Z. and Li R. (2020) Joint optimization of road classification and road capacity for urban freight transportation networks. Journal of Transportation Engineering, Part A: Systems, 146, 04020122.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

IMPACTO DA RECIDIVA NA SUPERVIVENCIA EN CÁNCER COLORRECTAL: ENFOQUE MULTIESTADO

Vanesa Balboa Barreiro[1,2], Sonia Pértega Díaz[1,2] , Teresa García Rodríguez[2], Cristina González Martín[1,2], Teresa Seoane Pillado[1,2]

[1] Universidade da Coruña, Rheumatology and Health Research Group, Department of Health Sciences, Faculty of Nursing and Podiatry, Ferrol, Spain

[2] Instituto de Investigación Biomédica de A Coruña (INIBIC), Nursing and Health Care Research Group, A Coruña, Spain.

RESUMO

O cancro é unha das enfermidades máis frecuentes a nivel mundial, encontrándose entre as principais causas de morbilidade e mortalidade, sendo o cancro colorrectal (CCR) a segunda causa de morte por tumor. Estudáronse 994 casos incidentes de CCR sometidos a resección con intención curativa co obxectivo de analizar o efecto de diferentes factores prognósticos e o impacto da recorrencia no prognóstico da enfermidade, mediante o desenrolo de modelos multiestado (MSM). Na análise multiestado, incluíuse a recorrencia como un estado intermedio na progresión da enfermidade. Ao final do seguimento, segundo as probabilidades de ocupación, entorno ao 50% dos pacientes estaban vivos e sen recorrencia, o 4,2% vivos con recorrencia e o 8,8% mortos por CCR. A recorrencia ten un impacto negativo no prognóstico do CCR. O estadio avanzado de cancro no diagnóstico de CCR identificouse como o factor prognóstico máis alto para as mortes sen recorrencia.

Palabras e frases chave: Colorectal Neoplasms; Neoplasm Recurrence; Prognosis

1. INTRODUCIÓN

O cancro é unha das enfermidades máis frecuentes a nivel mundial, considerado unha das principais causas de morbilidade e mortalidade, sendo o cancro colorrectal (CCR) o cuarto cancro máis diagnosticado e a quinta causa de morte por tumor. As taxas de supervivencia dos pacientes con CCR melloraron significativamente nas últimas décadas, superando na actualidade o 50% aos 5 anos dende o diagnóstico. Esta mellora na supervivencia pode deberse a melloras no diagnóstico, tratamento e técnicas cirúrxicas, así como ao estadio no momento do diagnóstico. A resección cirúrxica é o principal tratamento curativo para o 80% dos pacientes con CCR non metastáticos. Non obstante, máis do 40% dos pacientes elixibles para unha resección potencialmente curativa desenvolven enfermidades recorrentes durante o seguimento, presentando un maior risco de morte. Un mellor coñecemento do risco de recorrencia permitiría mellorar o seguimento dos pacientes sometidos a cirurxía curativa e, unha selección de terapias ou tratamentos máis axeitados en función do risco de recorrencia estimado do paciente.

O modelo de riscos proporcionais de Cox é o modelo máis empregado para a análise de supervivencia en epidemioloxía e, particularmente, en estudos de cancro. Algúns estudos de factores prognósticos de CCR utilizaron esta metodoloxía incluíndo a recorrencia como covariable dependente do tempo para avaliar o seu impacto sobre a mortalidade. Sen embargo, esta técnica asume que o efecto que pode ter a covariable de prognóstico basal sobre a mortalidade é o mesmo en pacientes con e sen recorrencia. Outra limitación desta metodoloxía é que analiza un único evento, polo tanto, non permite avaliar o efecto dos factores prognósticos basais sobre o risco de recorrencia. Os modelos multiestado (MSM) intentan superar estas limitacións, permitindo separar os efectos dos factores prognósticos en

diferentes eventos clínicos. A pesar da súa utilidade, son poucos os estudos recentes nos que se explora o seu uso para analizar o prognóstico dos pacientes con CCR e, como consecuencia, os resultados do emprego desta metodoloxía son escasos e limitados a series con un número reducido de pacientes nas que só se analizan algúns factores prognósticos.

O obxectivo deste estudo é investigar a utilidade dos MSM para analizar o impacto de diferentes factores prognósticos, na recorrencia do tumor e na morte específica do cancro, e para avaliar a importancia da recorrencia como evento intermedio para o prognóstico dos pacientes con CCR.

2. MATERIAL E MÉTODOS

Deseño e tipo de estudo

Estudo observacional de seguimento ambiespectivo de casos incidentes de CCR.

Criterios de inclusión: pacientes adultos (≥18 anos) con diagnóstico histopatolóxico de CCR (International Disease Classification 153-154), recrutados no Complexo Hospitalario Universitario A Coruña (CHUAC) entre 2006-2013.

Criterios de exclusión: casos prevalentes ou recorrentes, casos con múltiples cancros, casos tratados só en hospitais privados, casos detectados mediante cribado de CCR e casos diagnosticados noutro hospital.

Para dar resposta ó obxectivo deste estudo, só foron incluídos os pacientes sometidos a resección con intención curativa e que estaban libres de enfermidade despois da cirurxía e quimiorradioterapia adxuvante e neoadxuvante (N=994).

Parte desta cohorte foi incluída nun proxecto multicéntrico (estudos DECCIRE I e DECCIRE II(23,24)). Conta co consentimento informado dos pacientes e ca aprobación do Comité Ético de Investigación Clínica de Galicia (códigos 2004/159, 2009/160).

Recollida de datos

Os casos foron identificados a través do Departamento de Anatomía Patolóxica e, tras a obtención dos consentimentos informados, recolléronse os datos a través de entrevistas persoais, realizadas por enfermeiras adestradas e mediante a revisión das historias clínicas.

Medidas

Factores sociodemográficos, antecedentes familiares de cancro e comorbilidade (índice de Charlson). Variables relacionadas co CCR no momento do diagnóstico, incluíndo a localización do tumor, o grao histolóxico, o estadio TNM e o nivel de antíxeno carcinoembrionario (CEA).

Os eventos de interese foron a recorrencia e a morte. A recorrencia foi definida como recorrencia local e/ou metástase a distancia, considerando o evento que ocorrese en primeiro lugar. A morte clasificouse como morte relacionada co CCR ou non relacionada co cancro.

Análise estatística

Realizouse un análise descritivo das variables recollidas no estudo.

O tempo de seguimento foi calculado como o tempo dende a data da cirurxía ata a data de falecemento ou da última vez na que se tiña constancia de que o individuo estaba vivo. Estendeuse o seguimento ata novembro de 2019.

Utilizáronse MSM para separar os efectos dos factores prognósticos sobre o risco de recorrencia dos seus efectos sobre o risco de morte. Deseñouse un MSM cos seguintes estados: (1) estado inicial, "vivo sen recorrencia", (2) estado de transición "vivo con recorrencia" e dous estados absorbentes (3) "morte relacionada co CCR" e (4) "morte non relacionada co cancro" como evento competitivo (Figura 1). Para este estudo, as transicións de maior relevancia foron: transición de vivo e libre de recorrencia a recorrencia (1→2); de vivo e libre de recorrencia a morte relacionada co CCR (1→3) e, de recorrencia a morte relacionada co CCR (2→3). As probabilidades de ocupación e transición entre estados foron estimadas mediante o estimador Markoviano (Aalen-Johansen). Para examinar o efecto das covariables, calculouse o risco acumulativo empírico, mediante o modelo estratificado de riscos proporcionais de Cox en cada transición. Os coeficientes de regresión das covariables para cada transición estimáronse a través dun modelo de Cox específico para a transición. Este enfoque permítenos especificar riscos de transición basais separados para cada transición. O modelo semiparamétrico incluíu os seguintes factores prognósticos: idade no diagnóstico, sexo, estadio TNM, grao histolóxico, localización do tumor e CEA. Todas as análises estatísticas realizáronse mediante SPSS 24.0 e os paquetes *cmprsk*, *mstate* e *survival*, dispoñibles en R v3.5.1. Consideráronse estatísticamente significativos os p-valores bilaterais <0,05.

3. RESULTADOS

Foron estudiados un total de 994 pacientes con CCR, cun seguimento medio de 85,5±36,5 meses (mediana=91,5 meses) dende a cirurxía curativa. Do total de pacientes, 146 (14,7%) desenvolveron recorrencia local ou metástases tras o tratamento curativo (Figura 1), observándose unha mediana do tempo ata a recorrencia de 23,2 meses (IQR=16,7-38,2). Máis do 60% das recorrencias foron diagnosticadas entre os 12 e os 36 meses despois da cirurxía, cun aumento da incidencia acumulada dende o 1,6% aos 12 meses ata o 13,7% aos 60 meses tras a cirurxía. Ao final do seguimento faleceron 367 pacientes (36,9%), dos cales 92 presentaban unha recorrencia, supoñendo o 63% dos casos con recidiva. O 20,7% das mortes foron relacionadas co cancro, das cales, o 82,9% presentaran previamente recorrencia (Figura 1).

As probabilidades de ocupación mostran que o 93,0% dos pacientes seguían vivos e sen recaídas 1 ano despois da cirurxía, diminuíndo ata o 77,9% e 70,3% aos 3 e 5 anos, respectivamente. Neses mesmos tempos, o 1,6%, 8,8% e 8,4%, respectivamente, foron diagnosticados cunha recorrencia. A mortalidade relacionada co CCR a 5 anos aumenta do 3,8% ao 33,6% cando se comparan pacientes vivos e sen recorrencia un ano despois da cirurxía fronte a aqueles que tiveron unha recorrencia nese tempo, indicando un impacto negativo da recorrencia no prognóstico do CCR.

Ser diagnosticado cun estadio III-IV (HR=1,53; p=0,022) aumentou significativamente o risco de recorrencia. En canto á mortalidade relacionada co CCR, nos pacientes sen recidiva, a única variable que aumentou significativamente este risco foi a idade >=75 anos (HR=8,48; p=0,049). Despois dunha recorrencia, o risco de morte relacionada co CCR aumentou significativamente nos estadios III-IV (HR=2,35; p=0,004) e con puntuacións de comorbilidade máis altas (HR=1,54; p=0,003).

4. CONCLUSIÓNS

Este estudo demostra a utilidade do MSM para analizar en detalle o prognóstico, e os seus factores asociados, en pacientes libres de CCR despois da resección curativa.

Confirma o impacto negativo da recorrencia como evento intermedio na supervivencia, e analiza por separado a influencia de factores prognósticos simples sobre a recorrencia, a mortalidade relacionada co CCR e a mortalidade non relacionada co CCR.

REFERENCIAS

Aalen, O., Borgan, O., & Gjessing, H. (2008) Survival and event history analysis: a process point of view. Springer Science & Business Media.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012) Statistical models based on counting processes. Springer Science & Business Media.

Beyersmann, J., Wolkewitz, M., Allignol, A., Grambauer, N., & Schumacher, M. (2011) Application of multistate models in hospital epidemiology: advances and challenges. Biometrical Journal, 53(2), 332-350.

Compton, C. C., Tanabe, K. K., & Savarese, D. (2016) Pathology and prognostic determinants of colorectal cancer. UpToDate; Savarese, DM, Ed.; UpToDate: Waltham, MA, USA.

Cox, D. R. (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.

Dancourt, V., Quantin, C., Abrahamowicz, M., Binquet, C., Alioum, A., & Faivre, J. (2004) Modeling recurrence in colorectal cancer. Journal of Clinical Epidemiology, 57(3), 243-251.

De Angelis, R., Sant, M., Coleman, M. P., Francisci, S., Baili, P., Pierannunzio, D., ... & Capocaccia, R. (2014) Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE-5—a population-based study. The Lancet Oncology, 15(1), 23-34.

De Wreede, L. C., Fiocco, M., & Putter, H. (2010) The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. Computer Methods and Programs in Biomedicine, 99(3), 261-274.

Gilard-Pioc, S., Abrahamowicz, M., Mahboubi, A., Bouvier, A. M., Dejardin, O., Huszti, E., ... & Quantin, C. (2015) Multi-state relative survival modelling of colorectal cancer progression and mortality. Cancer Epidemiology, 39(3), 447-455.

Gray, R. J. (1988) A class of k-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics, 1141-1154.

Hougaard, P. (1999) Multi-state models: a review. Lifetime Data Analysis, 5(3), 239-264.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., & Andersen, P. K. (2009) Multi-state models for the analysis of time-to-event data. Statistical Methods in Medical Research, 18(2), 195-222.

Meira-Machado, L., de Uña-Álvarez, J., & Cadarso-Suárez, C. (2006) Nonparametric estimation of transition probabilities in a non-Markov illness–death model. Lifetime Data Analysis, 12, 325-344.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209-249.

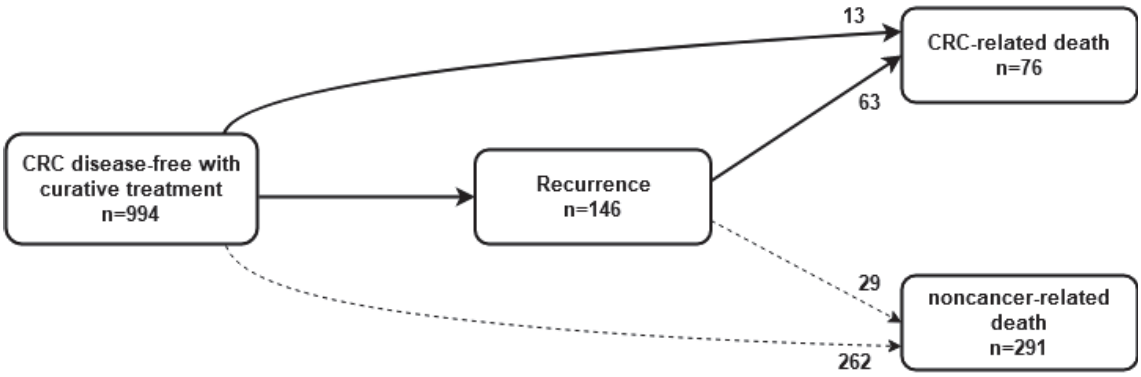


Figura 1: Modelo multiestado de recorrencia e mortalidade en cancro colorrectal

Comparative Evaluation of Survival Analysis Models: Cox Proportional Hazards versus Random Survival Forest Using Real-World Data

Ana Borges[1] e Mariana Carvalho[2]

[1] CIICESI, ESTG, Politécnico do Porto
[2] CIICESI, ESTG, Politécnico do Porto

ABSTRACT

Survival analysis is a crucial statistical tool used in various fields, including medicine, finance, and engineering, to analyse time-to-event data. In this study, we aimed to compare the performance of three distinct survival analysis models: the Cox Proportional Hazards (PH) model, Random Forest Survival Analysis (RFSA) with axes-based splitting and Accelerated Oblique Survival Random Forest (AOSRF). Our objective is to assess the predictive accuracy of these models using real-world data, with a focus on metrics such as the Area Under the Curve (AUC), Brier score, and Integrated Prediction Accuracy (IPA). Our preliminary results indicate that the choice of survival analysis model significantly impacts predictive performance. While the Cox PH model demonstrated interpretability, it struggled with capturing complex non-linear relationships. RFSA with axes-based splitting and AOSRF exhibited promising results, outperforming the Cox PH model in terms of AUC and Brier score.

In conclusion, our study contributes to the understanding of survival analysis modelling by comparing the Cox PH model with machine learning-based approaches like RFSA and AOSRF. We shed light on the trade-offs between interpretability and predictive accuracy by utilising real-world data and evaluating multiple metrics.

Keywords: survival analysis; Cox Proportional Hazards Model, axes-based Random Survival Forest, oblique Random Survival Forest.

ACKNOWLEDGEMENT

This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through project UIDB/04728/2020.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Disentangling Hospitalisation Trajectories

Rita Gaio[1,2], Daniel Cordeiro[1], Bárbara Peleteiro[3,4,5], Lucybel Moreira[3], Elsa Guimarães[3], Raquel Cadilhe[3], Ana Azevedo[3,4,5]

[1] Department of Mathematics, Faculty of Sciences, University of Porto
[2] Centre of Mathematics of the University of Porto
[3] Centre of Hospital Epidemiology, Centro Hospitalar Universitário de São João
[4] Department of Forensic and Public Health Sciences, and Medical Education; Faculty of Medicine, University of Porto
[5] Public Health Institute, University of Porto

ABSTRACT

The efficient management of patient flows in healthcare settings is a complex and critical issue. This work presents an inventive approach to analyse pediatric hospitalisation trajectories at Centro Hospitalar Universitário de São João, a portuguese public teaching tertiary care hospital in the city of Porto.

The analysis comprises data from 3133 hospitalisation episodes of newborns, infants, children and adolescents, aged between 0 months and 18 years old, from December 2021 to November 2022, excluding newborns who were discharged within 48 hours of birth. Each hospitalisation trajectory is represented by a sequence of a maximum of nine paediatric services and the associated lengths of stay in each of them.

Based on the specifics of the data, a customised dissimilarity is developed, capturing discrepancies between the sequences of pairs of services and lengths of stay. The novelty of the methodology lies in combining service dissimilarities with time-based distances, ensuring that the order of the pairs in each trajectory is not disregarded. More precisely, the dissimilarity consists of the sum of three terms: (1) the first term includes a comparison between services, which excludes common items and takes into account dissimilarities specifically defined by the medical team; (2) the second term regards the lengths of stay, distinguishing those arising from the same services in the two sequences being compared, from those of different services; (3) the last term makes use of the Needleman algorithm of sequence alignment in Bioinformatics, ensuring that two sequences consisting of the same pairs but in a different order are not totally equal.

From the described dissimilarity, seven homogeneous clusters are then defined based on the partition around medoids clustering algorithm. Five clusters mostly consist of trajectories through a single service while two clusters mostly include trajectories through two different services. In essence, the clustering procedure is isolating some of the trajectories passing through a single paediatric service while forming different groups for those consisting of multiple services.

The present approach provides a comprehensive understanding of hospitalisation trajectories, resulting in meaningful insights for hospital administrators, healthcare providers, and policy makers, thus enabling them to improve healthcare delivery at the paediatric patient care level.

Keywords: Partition Around Medoids Algorithm, Needleman Algorithm, Dissimilarity Matrix.

Acknowledgement: Rita Gaio was partially supported by CMUP, member of LASI, which is financed by national funds through FCT ? Fundação para a Ciência e a Tecnologia, I.P., under the projects with reference UIDB/00144/2020 and UIDP/00144/2020

REFERENCES

Kaufman, L. and Rousseeuw, P. J. (2009) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, vol. 48, no. 3, 443–453.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

EMPIRICAL SOCIAL SCIENCES AND GLOBAL SPATIAL TRENDS

Fernando Bruna[1], Román Mínguez[2]

[1] Universidade da Coruña, C+D Research Group, Department of Economics, School of Economics, A Coruña, Spain. f.bruna@udc.es

[2] Department of Applied Economics I, University of Castilla-La Mancha, Cuenca, Spain. roman.minguez@uclm.es

ABSTRACT

Social sciences, particularly economics, are interested on causality. Theories usually predict the mean behaviour of a variable, which implies a prediction of the spatial distribution of the variable when working with areal data. That made, however, to ignore the possible presence of global spatial trends in the data, which has strong consequences. This paper tries to provoke a dialog between spatial statistics, geostatistics and social sciences, with special attention to the analysis of causal relationships.

Keywords: Non-stationarity, spatial dependence, autocorrelation, spurious regression

1. INTRODUCTION

Stationarity assumptions imply constant mean and variance on the space. The simplest form of departure from spatial stationarity is to allow the mean to depend on location. Such varying mean is a spatial trend. When the mean of a spatial series has a systematic trend across the geographical coordinates of the data points, the series are said to be non-stationary of first order. That is associated to the presence of unit roots in the data and unreliable statistical tests designed for stationary variables. Spatial econometrics has shown, however, low interest about unit roots in spatial data. Exceptions are Mur and Trívez (2003), Lauridsen and Kosfeld (2007) or Müller and Watson (2023).

Non-stationarity of first order has strong implications. It may produce spurious regressions, as in time series (Granger & Newbold, 1974). Global spatial trends may also produce misleading detection of (local) spatial dependence. If, for instance, the mean is smoothly changing from North to South, values for close neighbors will be correlated. That does not necessarily imply substantive interactions between close territories.

2. REGIONAL SCIENCE, IN THE MIDDLE OF TRADITIONS THAT ARE NOT DEBATING

Despite their different traditions, both geostatistics and spatial statistics share similar approach regarding non-stationarity, which may be called “global-to-local”: first model the possible presence of global trends, then, model local interactions. “Trend surface analysis” is part of their standard toolbox: mapping the expected values of data in a continuous grid of spatial points.

Conversely, spatial econometrics evolved as an extension to space of autocorrelated time series but keeping a focus on the search of causal relationships. Even Mur and Trívez (2003), who alerted about the risks of spatial trends, followed a “local-to-global” approach, from a spatial regressive process to a global trend. Spatial econometrics, as a discipline, did not open the Pandora box of spatial trends. A few authors propose to add a trend component in spatial regression models (McMillen, 2012; Basile et al., 2021). Controlling out a possible trend is not a good idea, however, if the research goal is to test a theory predicting the behavior of the mean of a variable. But ignoring it, may be misleading (Bruna, 2023).

Regional science has also received other fertilizations. John Stewart, the astrophysicist that in 1947 applied the law of gravity to demographics, used a variant of trend surface analysis to plot contour maps. Later economic geographers followed Stewart’s legacy to display the change of the strength of “Market

Potential" through space. With some exceptions, such as Carruthers and Mulligan (2012), trend surface analysis is currently ignored.

3. EXAMPLE: SPATIAL TREND IN EUROPEAN REGIONAL INCOME PER CAPITA

Figure 1 displays a choropleth map of an indicator of income per capita (GVApc) for a cross-section of regions. Values are divided into eight quantiles, with darker colors for higher values. The spatial distribution appears to have a rough core-periphery pattern. Only a few regions with high log GVApc seem to be located outside the so-called "blue banana",[1] notably regions in Nordic countries. Figure 1 also includes two scatterplots of log GVApc against each of the spatial coordinates of regional centroids, (E, N), which are the regional geographical centers. Horizontal axes represent location measured in kilometers from an origin in the South-East. Dashed lines are regression lines, while solid lines result from a locally weighted smoother. The solid line shows a core-periphery spatial pattern in the West-to-East direction, with higher income for distances around the geographical center of Europe. The right plot shows an increasing trend from South to North, combined with some core-periphery tendency.

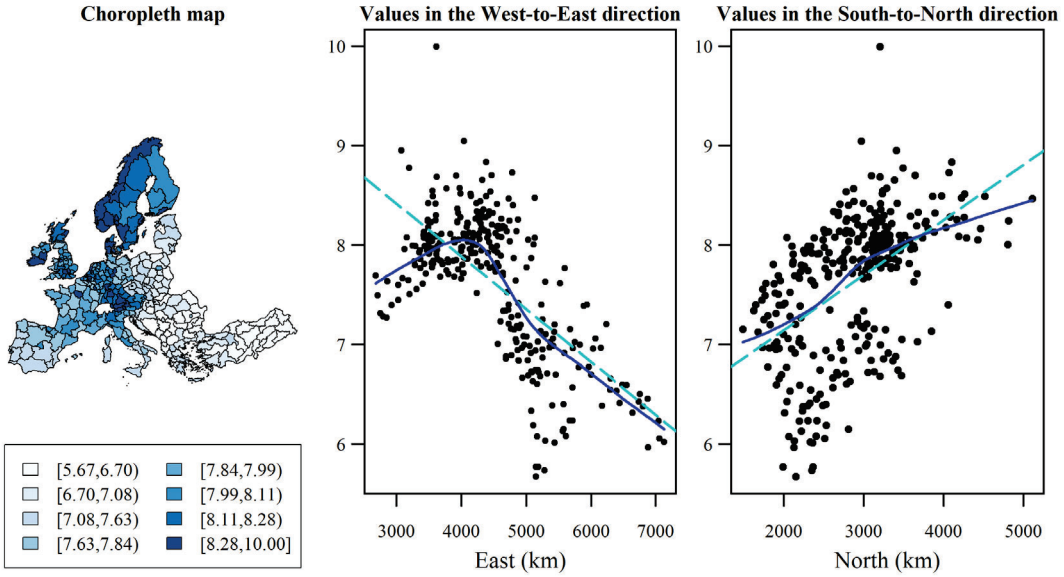


Figure 1: Logarithm of Gross Value Added per capita (311 regions, 2019)

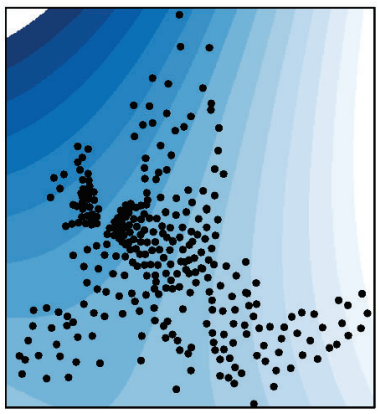


Figure 2: Predictions on a grid from a polynomial trend surface of order 2

[1] See https://en.wikipedia.org/wiki/Blue_Banana.

Figure 2 shows a level plot of a trend surface created after estimating regional values as a quadratic function of the $(E,\ N)$ centroids' coordinates, such as the following: $\log GVApc = C + \beta_1 E + \beta_2 N + \beta_3 E^2 + \beta_4 N^2 + \beta_5 EN$. Estimation results are used to predict values on a grid of 124 thousand points of spatial coordinates. Similar predictions are plotted as colored bands, except for white out-of-range predictions. Black points indicate the coordinates of the regional centroids, as references of the underlying map. Colors at those points show that income per capita displays a systematic change from North-West to South-East. It also displays a core-periphery pattern, with higher inland predictions in the "blue banana".

4. CONCLUSIONS

Bruna (2023) shows that ignoring global core-periphery spatial trends in the data led to misleading interpretations in previous research testing the New Economic Geography (NEG) for European regions. Those spatial trends were also affecting the instruments used in instrumental variables estimations.

Literature on spatial regimes (clubs) or local models offers alternatives but does not necessarily solve problems derived from the presence of spatial trends. Empirical researchers should be much more careful about the spatial properties of the data.

REFERENCES

Basile, R., De Benedictis, L., Durban, M., Faggian, A., & Mínguez, R. (2021). The impact of immigration on the internal mobility of natives and foreign-born residents: Evidence from Italy. *Spatial Economic Analysis*, *16*(1), 9–26. https://doi.org/10.1080/17421772.2020.1729997

Bruna, F. (2023). Market Potential, spatial theories, and spatial trends. *Spatial Economic Analysis*, *(first revision submitted)*.

Carruthers, J., & Mulligan, G. F. (2012). The plane of living and the precrisis evolution of housing values in the USA. *Journal of Economic Geography*, *12*(4), 739–773. https://doi.org/10.1093/jeg/lbr045

Lauridsen, J., & Kosfeld, R. (2007). Spatial cointegration and heteroscedasticity. *Journal of Geographical Systems*, *9*(3), 253–265. https://doi.org/10.1007/s10109-007-0048-y

McMillen, D. P. (2012). Perspectives on Spatial Econometrics: Linear Smoothing with Structured Models. *Journal of Regional Science*, *52*(2), 192–209. https://doi.org/10.1111/j.1467-9787.2011.00746.x

Müller, U. K., & Watson, M. W. (2023). *Spatial Unit Roots*. Department of Economics, Princeton University. https://www.princeton.edu/~umueller/

Mur, J., & Trívez, F. J. (2003). Unit Roots and Deterministic Trends in Spatial Econometric Models. *International Regional Science Review*, *26*(3), 289–312. https://doi.org/10.1177/0160017603253790

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

A CORRESPONDENCIA ENTRE PIERRE DE FERMAT E BLAISE PASCAL

José Nicanor Alonso Álvarez[1], Miguel Ángel Mirás Calvo[2] e Carmen Quinteiro Sandomingo[3]

[1]Departamento de Matemáticas, Universidade de Vigo
[2]Departamento de Matemáticas, Universidade de Vigo
[3]Departamento de Matemáticas, Universidade de Vigo

RESUMO

Como testemuñan innumerables restos arqueolóxicos, a afección dos seres humanos polos xogos de azar é unha constante ao longo da historia. Xogos, naturalmente, onde ademais do pracer de xogar adoitaba existir unha recompensa, as máis das veces económica, para aquel que ganaba e que, por tanto, contaban cunhas regras para o seu desenvolvemento. Pero, que facer se, por circunstancias alleas aos xogadores, estes deben abandonar unha partida unha vez iniciada sen podela retomar máis adiante e sen que ninguén ganase aínda? Como realizamos nese caso un repartimento xusto da cantidade apostada?

Este problema, estudado primeiramente por autores italianos do século XVI con resultados pouco satisfactorios, foille presentado no século XVII a un dos máis grandes científicos franceses, Blaise Pascal. Pero Pascal non se limitou a resolvelo, senón que ademais compartiu as súas novidosas ideas con outro xenio das matemáticas, Pierre de Fermat, que acadou outro método axeitado para solucionalo. Da correspondencia ente ambos autores xurdiría a teoría da probabilidade, unha rama das matemáticas que as relacionaba con algo tan aparentemente oposto a elas como é o azar.

Neste traballo presentamos por vez primeira a tradución ao galego, dende o orixinal en francés, das oito cartas que se conservan da correspondencia na que, durante o ano 1654, Fermat e Pascal discuten, entre outros, o problema do repartimento. Asemade, incluímos uns breves apuntamentos biográficos e abundantes notas explicativas. Nelas concretamos aspectos matemáticos, que para unha mellor comprensión detallamos en notación moderna, pois naquela época non existían a meirande parte dos signos que usamos na actualidade ou eran de uso infrecuente. Esperamos que o texto teña tamén un interese histórico e mesmo literario, pois permite ver como se facía ciencia hai máis de tres séculos, así como as fórmulas de tratamento e cortesía de uso obrigado entre os científicos daquela época.

Keywords: Blaise Pascal, Pierre de Fermat, historia da probabilidade, o problema da aposta vantaxosa, o problema do repartimento.

REFERENCIAS

Basulto Santos, J. e Camúñez Ruiz, J. A. (2007). La geometría del azar. Editorial Nivola.

David, F. N. (1962). Games, gods and gambling. Hafner Publishing Company, New York.

Diaconis, P. e Skyrms, B. (2018). Ten great ideas about chance. Princeton University Press.

Rueda, R. (2018). Blaise Pascal y Pierre de Fermat ¿Los fundadores de la probabilidad? Miscelánea Matemática, 65:55–68.

Sandford, V. e Merrington, M. (1998). Fermat and Pascal on probability. https://www.york.ac.uk/depts/maths/histstat/pascal.pdf.

Tannery, P. e Henry, C., editores (1894). Oeuvres de Fermat. Tome deuxième. Correspondance. Gauthiers-Villars et fils.

Tetra STATIS-Dual. STATIS para datos binarios.

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

[1] Departamento de Estadística, Universidad de Salamanca

RESUMEN

Cuando tenemos un conjunto de varias tablas de datos y deseamos extraer su estructura común, podemos utilizar una técnica llamada STATIS-ACT (Lavit et al., 1994; des Plantes, 1976). Cuando las tablas comparten un conjunto común de individuos, utilizamos la versión regular para obtener un mapa euclidiano de consenso de ellos. Cuando tenemos variables comunes, utilizamos la versión dual para obtener una estructura de correlación de consenso (Abdi et al., 2012).

Para datos binarios, la versión regular se llama DISTATIS (Abdi et al., 2005) y encuentra una configuración común de los individuos basada en matrices de distancia, pero no existe una versión dual para este tipo de datos. En este trabajo proponemos una versión dual de STATIS-ACT basada en correlaciones tetracóricas y desarrollamos las adaptaciones necesarias para el método. Hemos denominado a esta propuesta Tetra STATIS-Dual. También incluiremos las representaciones gráficas asociadas al método.

Ilustraremos la técnica utilizando un conjunto de datos reales.

Palabras y frases clave: STATIS, STATIS Dual, datos binarios

REFERENCIAS

Abdi, H., Williams, L. J., Valentin, D., and Bennani-Dosse, M. (2012). STATIS and DISTATIS: optimum multitable Principal Component Analysis and three way metric multidimensional scaling. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2):124–167.

Abdi, H., O'Toole, A. J., Valentin, D., and Edelman, B. (2005). DISTATIS: The analysis of Multiple Distance Matrices. En 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, pp. 42–42. IEEE.

des Plantes, H. L. (1976). Structuration des tableaux à trois indices de la statistique. Théorie et application d'une méthode d'analyse conjointe. Tesis doctoral, Université des sciences et techniques du Languedoc.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The ACT (STATIS method). Computational Statistics y Data Analysis, 18(1):97–119.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

HELPING SPHERICAL NONPARAMETRIC DENSITY ESTIMATION WITH A PARAMETRIC HINT

María Alonso-Pena[1], Gerda Claeskens[1,3] and Irène Gijbels[2,3]

[1]ORSTAT, KU Leuven
[2]Department of Mathematics, KU Leuven
[3]Leuven Statistics Research Center (LStat), KU Leuven

ABSTRACT

In this work, we consider density estimation of hyperspherical variables, which are observations defined on the surface of a sphere with arbitrary dimension. Although kernel estimators for this type of variables have already been studied (Hall et al., 1987; Bai et al., 1988), we aim to improve the properties of the classical hyperspherical kernel density estimator by providing it with a parametric hint or parametric guide, following the ideas of Hjort and Glad (1995). Essentially, a parametric model is estimated and, subsequently, the misspecified quantity is estimated nonparametrically. This leads to a smaller bias than the classical kernel estimator in most cases, while the variance of the estimator remains the same. This methodology has a huge advantage when the parametric guide contains the uniform distribution as a particular case, as it is the case with the well-known von Mises-Fisher distribution. We derive the asymptotic bias and variance of the estimator and prove its asymptotic normality. In addition, we perform a computational study showing that, when the von Mises-Fisher guide is employed, the parametrically guided estimator only improves the classical estimator, even in cases where the guide model is very far from the true density. Lastly, we also provide some real data examples to illustrate the use of the new estimator.

Keywords: Density estimation; Directional data; Kernel smoothing; Nonparametric statistics; Parametric guide; Spherical data.

REFERENCES

Bai, Z. D., Rao, C. R. and Zhao, L. C. (1988) Kernel estimators of density function of directional data. Journal of Multivariate Analysis, 27, 24–39.

Hall, P.,Watson, G. S. and Cabrera, J. (1987) Kernel density estimation with spherical data. Biometrika, 74, 751–762.

Hjort, N. L. and Glad, I. K. (1995) Nonparametric density estimation with a parametric start. The Annals of Statistics, 23, 882–904.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

NONPARAMETRIC ESTIMATION OF THE SPARSITY FUNCTION

Mercedes Conde-Amboage[1,2] and César A. Sánchez-Sellero[1,2]

[1] Department of Statistics, Mathematical Analysis and Optimization. Faculty of Mathematics. Universidade de Santiago de Compostela (USC).
[2] CITMAga, 15782 Santiago de Compostela, Spain.

ABSTRACT

It is perfectly natural that the precision of quantile estimates should depend on the inverse of the density evaluated at the quantile, called sparsity function, because it reflects the density of observations near the quantile of interest. If the data are very sparse at the quantile of interest, this quantile will be difficult to estimate, but when the density is high, the quantile is more precisely estimated.

Moreover, the asymptotic distribution associated with the parametric quantile regression estimator also depends on the inverse of the conditional density function evaluated at the quantile of interest. In the regression context, this function plays an analogous role to the variance of the errors in least squares estimation of classical mean regression.

Along this talk a new nonparametric estimator of the conditional sparsity function, based on kernel ideas, will be presented. Furthermore, different bandwidth selectors will be compared using a Monte Carlo simulation study.

Keywords: quantile regression; nonparametric estimation; sparsity function..

Estimación non paramétrica de rexións de alta densidade para datos direccionais

Diego Bolón[1,2], Rosa M. Crujeiras[1,2] e Alberto Rodríguez Casal[1,2]

[1] CITMAga, Santiago de Compostela, España.
[2] Departamento de Estatística, Análise Matemática e Optimación, Universidade de Santiago de Compostela, España.

RESUMO

As rexións de alta densidade (HDRs polas súas iniciais en inglés) son os conxuntos onde a función de densidade da variable aleatoria supera un determinado valor prefixado de antemán. A estimación destes conxuntos resulta ser unha ferramenta moi útil para a visualización e exploración de datos. Por exemplo, esta técnica resultou ser de utilidade para aproximar a localización de campos de minas a partir de observacións aéreas, a análise de datos seismolóxicos (Huo e Lu, 2004) e a detección de *outliers* dentro dunha mostra (Markou e Singh, 2003).

A estimación de HDRs para datos euclídeos (uni ou multidimensionais) foi tratada ampliamente na literatura estatística. Porén, a estimación de HDRs noutros contextos, como poden ser os datos circulares e direccionais, non recibiu especial atención ata hai pouco tempo, véxase Saavedra-Nieves e Crujeiras (2022) e Cholaquidis et al. (2022).

Tendo en conta o anterior, neste traballo introdúcese unha nova técnica non paramétrica para a estimación de HDRs direccionais baixo certas condicións de regularidade. Máis concretamente, este novo método de estimación consiste nunha adaptación do estimador proposto en Walther (1997) a datos direccionais, que combina información xeométrica xunto cun estimador non paramétrico da función da densidade.

Palabras e frases chave: Estimación non paramétrica, estimación de conxuntos, conxuntos de alta densidade, datos direccionais.

REFERENCIAS

Cholaquidis, A., Fraiman, R., and Moreno, L. (2022) Level set and density estimation on manifolds. Journal of Multivariate Analysis, 189, 104925.

Huo, X. and Lu, J.-C. (2004) A network flow approach in finding maximum likelihood estimate of high concentration regions. Computational Statistics & Data Analysis, 46, 33–56

Markou, M. and Singh, S. (2003) Novelty detection: A review–part 1: Statistical approaches. Signal Processing, 83, 2481–2497.

Saavedra-Nieves, P., Crujeiras, R.M. (2022) Nonparametric estimation of directional highest density regions. Advances in Data Analysis and Classification, 16, 761–796.

Walther, G. (1997) Granulometric smoothing. Annals of Mathematical Statistics, 25, 2273–2299.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

ESTIMATING THE SHAPE FUNCTIONS

Juan Carlos Pardo-Fernández[1] and María Dolores Jiménez-Gamero[2]

[1]Centro de Investigación e Tecnoloxía Matemática de Galicia, Universidade de Vigo
[2]Departamento de Estadística e Investigación Operativa, Universidad de Sevilla

ABSTRACT

Arriaza *et al.* (2019) introduced the right and left shape functions, which enjoy interesting properties in terms of describing the global form of a distribution. In this talk, we will propose nonparametric estimators of those functions. The estimators involve nonparametric estimation of the quantile and density functions. Pointwise and uniform consistency are proved under general regularity assumptions, as well as the limit in law. Simulations are included to study the practical performance of the proposed estimators. The analysis of a real data set illustrates the methodology.

Keywords: Shape functions; nonparametric estimation; pointwise convergence; weak convergence.

REFERENCES

Arriaza, A., Di Crescenzo, A., Sordo, M.A. and Suárez-Llorens, A. (2019) Shape measures based on the convex transform order. Metrika 82, 99-124.

Pósters

Problemas de asignación de costes en autopistas con usuarios agrupados

Marcos Gómez–Rodríguez[1], Laura Davila–Pena[2,3] y Balbina Casas–Méndez[4]

[1]Máster en Técnicas Estadísticas, Facultad de Informática, Universidad de A Coruña, Campus de Elviña, 15071 A Coruña.

[2]Grupo de investigación MODESTYA, Departamento de Estadística, Análisis Matemático y Optimización, Facultad de Matemáticas, Universidad de Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela.

[3]Department of Analytics, Operations and Systems, Kent Business School, Kent University, Canterbury Campus, Kent, England, CT2 7NZ.

[4]CITMAga, Grupo de investigación MODESTYA, Departamento de Estadística, Análisis Matemático y Optimización, Facultad de Matemáticas, Universidad de Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela.

RESUMEN

Una de las aplicaciones de los juegos cooperativos con utilidad transferible es la determinación de las tarifas que deben pagar los usuarios de una determinada instalación, cuyos costes de construcción o mantenimiento deben de ser recuperados y donde ciertos criterios de eficiencia y equidad inspiran las soluciones consideradas (Fiestras-Janeiro et al., 2011). Uno de los trabajos más conocidos en este campo es debido a Littlechild y Owen (1973), donde se definen los juegos del aeropuerto y los costes de construcción de una pista de aterrizaje son asignados entre los aviones de las compañías aéreas que la utilizan. Hu et al. (2012) estudian esta situación por medio de reglas empleadas habitualmente en el análisis de problemas de bancarrota. Por otro lado, el uso de la teoría de juegos cooperativos en la búsqueda de un peaje óptimo para las autopistas se ha investigado en varios trabajos siguiendo distintos enfoques, entre otros en Villarreal-Cavazos y García-Díaz (1985), Castaño Pardo y García-Díaz (1995), Dong et al. (2012), Kuipers et al. (2013), Sudhölter y Zarzuelo (2017) y, más recientemente, en Wu et al. (2022).

En este trabajo analizamos cómo distribuir los costes fijos de una autopista entre sus usuarios a través de peajes, considerando distintas clases o grupos de vehículos que utilizan la autopista. Para ello, se hace uso de los juegos generalizados de autopistas, introducidos en el trabajo de Sudhölter y Zarzuelo (2017) junto con un sistema de uniones a priori que representan a grupos con mayor poder de negociación, capaces de obtener reducciones en sus tarifas. Este modelo viene motivado por el hecho de que, recientemente, la autopista AP-9 de España (Figura 1) empezó a ofrecer tarifas especiales para determinados grupos, como trabajadores, camioneros o personas que realizan más de 20 viajes al mes.

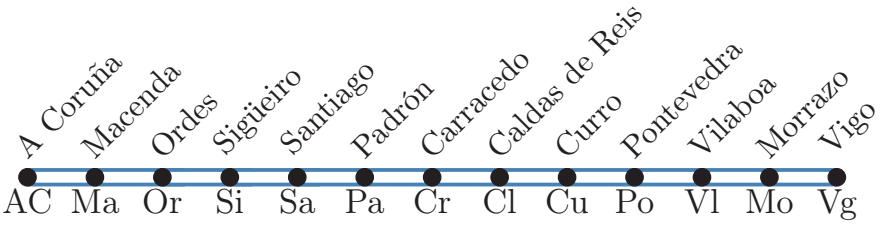


Figura 1: Tramos de la autopista AP-9 entre A Coruña y Vigo.

En particular, se obtienen resultados relativos a la obtención de expresiones sencillas y caracterizaciones axiomáticas para el valor de Owen (Owen, 1977) y el valor de Tijs

coalicional (Casas-Méndez et al., 2003), en el contexto de los problemas generalizados de autopistas con uniones a priori. Además, se propone y estudia un nuevo valor, el valor de Shapley-Tijs, que conjuga propiedades razonables de los dos anteriores y de modo que logra un reparto equitativo dentro de las uniones y, además, garantiza que alianzas establecidas entre diferentes uniones sean siempre beneficiosas para ellas.

En cuanto a la aplicabilidad de estos valores, cabe mencionar que en Kuipers et al. (2013), el valor de Shapley (Shapley, 1953) y el nucleolo (Schmeidler, 1969) se aplican al reparto de los costes de la autopista AP-68 de España, entre sus usuarios. No obstante, el estudio se limita únicamente a distribuir los costes entre los vehículos ligeros. La metodología introducida en nuestro trabajo la aplicamos a un caso real de datos de tráfico en la autopista AP-9 (AUDASA, s.f., Ministerio de Transportes, Movilidad y Agenda Urbana de España, 2019a y 2019b). El modelo que se adoptará en nuestro caso, para incluir las distintas clases de vehículos, se basa en el utilizado por Fragnelli et al. (2000) para analizar un problema de asignación de costes, denominado problema de costes de infraestructura, problema surgido durante la reorganización ferroviaria llevada a cabo en Europa en los años 90.

Para finalizar el resumen, consideramos que como futuro trabajo de investigación, podría ser interesante incorporar los costes de mantenimiento a nuestro modelo, además de los costes fijos, en la línea de los trabajos de Fragnelli et al. (2000) y Costa (2015).

Palabras y frases clave: Teoría de juegos, Problemas de autopistas generalizados, Asignación de costes, Uniones a priori, Valores coalicionales, Autopista AP-9.

AGRADECIMIENTOS

Este estudio es apoyado por la ayuda de I+D+I correspondiente al proyecto PID2021-124030NB-C32, financiado por MCIN/AEI/10.13039/ 501100011033/ y por "*ERDF A way of making Europe*"/EU. La investigación también ha sido financiada por el Grupo de Referencia Competitiva ED431C 2021/24 desde la Consellería de Cultura, Educación e Universidades, Xunta de Galicia.

REFERENCIAS

AUDASA (s.f.). Tarifas y descuentos. Disponible en: https://www.audasa.es/la-autopista/ tarifas-y-descuentos/#tab-id-3. Último acceso: 19 de Abril de 2023.

Casas-Méndez, B., García-Jurado, I., van den Nouweland, A., and Vázquez-Brage, M. (2003) An extension of the τ-value to games with coalition structures. European Journal of Operational Research, 148, 494–513.

Castaño Pardo, A. and García-Díaz, A. (1995) Highway cost allocation: An application of the theory of nonatomic games. Transportation Research Part A: Policy and Practice, 29, 187–203.

Costa, J. (2015) Valores Coalicionales en Juegos Cooperativos con Utilidad Transferible. Tesis Doctoral, Universidad de A Coruña. Disponible en: http://dm.udc.es/matematicas/sites/default/files/CostaBouzas_Julian_TD_2015.pdf. Último acceso: 19 de Abril de 2023.

Dong, B., Guo, G., and Wang, Y. (2012) Highway toll pricing. European Journal of Operational Research, 220, 744–751.

Fiestras-Janeiro, M. G., García-Jurado, I., and Mosquera, M. A. (2011) Cooperative games and cost allocation problems. Top, 19, 1–22.

Fragnelli, V., García-Jurado, I., Norde, H., Patrone, F., and Tijs, S. (2000) How to share railways infrastructure costs? In: F. Patrone, I. García-Jurado, and S. Tijs (Eds.), Game Practice: Contributions from Applied Game Theory (pp. 91–101). Boston, MA: Springer.

Hu, C., Tsay, M., and Yeh, C. (2012) Axiomatic and strategic justifications for the constrained equal benefits rule in the airport problem. Games and Economic Behavior, 75, 185–197.

Kuipers, J., Mosquera, M. A., and Zarzuelo, J. M. (2013) Sharing costs in highways: A game theoretic approach. European Journal of Operational Research, 228, 158–168.

Littlechild, S. C. and Owen, G. (1973) A simple expression for the Shapley value in a special case. Management Science, 20, 370–372.

Ministerio de Transportes, Movilidad y Agenda Urbana de España (2019a). Autopista AP-9, Ferrol - Frontera portuguesa. Disponible en: https://www.mitma.gob.es/recursos mfom/autopista ap-9 ferrol - frontera portuguesa 2022.pdf. Último acceso: 19 de Abril de 2023.

Ministerio de Transportes, Movilidad y Agenda Urbana de España (2019b). Mapa de tráfico de la DGC. Disponible en: https://mapas.fomento.gob.es/mapatraco/2019/. Último acceso: 19 de Abril de 2023.

Owen, G. (1977) Values of games with a priori unions. In: R. Henn, and O. Moeschlin (Eds.), Mathematical Economics and Game Theory (pp. 76–88). Berlin, Heidelberg: Springer.

Schmeidler, D. (1969) The nucleolus of a characteristic function game. SIAM Journal on Applied Mathematics, 17, 1163–1170.

Shapley, L. S. (1953) A value for n-person games. In: H. W. Kuhn, and A. W. Tucker (Eds.), Contributions to the Theory of Games (AM-28) (pp. 307–318). Princeton, NJ: Princeton University Press volume II.

Sudhölter, P. and Zarzuelo, J. M. (2017) Characterizations of highway toll pricing methods. European Journal of Operational Research, 260, 161–170.

Villarreal-Cavazos, A. and García-Díaz, A. (1985) Development and application of new highway cost-allocation procedures. Transportation Research Record, 1009, 34–41.

Wu, H., van den Brink, R., and Estévez-Fernández, A. (2022) Highway toll allocation. Tinbergen Institute Discussion Papers, 22-036/II, Tinbergen Institute. The Netherlands.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

TUGlabR, un paquete de R para xogos coalicionais

Iago Núñez Lugilde[1], Miguel Ángel Mirás Calvo[2],
Carmen Quinteiro Sandomingo[3] e Estela Sánchez Rodríguez[1].

[1] CINBIO, Universidade de Vigo, SiDOR. Departamento de Estatística e Investigación Operativa.
[2] Universidade de Vigo, RGEAF, Departamento de Matemáticas.
[3] Universidade de Vigo, Departamento de Matemáticas.

RESUMO

A teoría de xogos cooperativos modela problemas de toma de decisións nas que varios axentes, denominados xogadores, poden obter certos beneficios ao cooperar entre si para acadaren un resultado óptimo para todos. Desde a publicación do libro *The Theory of Games and Economic Behavior* [14], esta disciplina foi medrando e desenvolvéronse diversos modelos e aplicacións tanto para comprender as accións dos xogadores como para propor repartimentos ou solucións. [6] é un libro de referencia que cobre gran parte dos modelos e das solucións clásicas. Actualmente mesmo é posible atopar paquetes creados con **R** dedicados aos xogos coalicionais. Por exemplo, `GameTheory` [1], `CoopGame` [13] ou `ClaimsProblems` [11], dos cales, o último está exclusivamente dedicado ao cálculo do repartimento de recursos escasos, unha clase de problemas aplicable a situacións de bancarrota, á repartición de impostos ou á distribución de emisións de CO_2.

O proxecto `TUGlab` (Transferable Utility Games laboratory, [9, 10]) nace no ano 2006, tratando fundamentalmente de salientar os aspectos xeométricos da teoría de xogos cooperativos para 3 e 4 xogadores, sen preocuparse da complexidade matemática dos cálculos. Máis adiante, comeza o desenvolvemento de `TUGlabExtended` [7], onde xa se poden atopar funcións aplicables a xogos con calquera número de xogadores. Posteriormente, preséntase `TUGlabWeb` (`http://TUGlabweb.uvigo.es/TUGlabWEB2/index.php`, [3]), unha plataforma en liña na que se implementan as funcións básicas de `TUGlab`, de maneira que as persoas usuarias poden experimentar con xogos cooperativos soamente introducindo a función característica do xogo a analizar. Esta plataforma está sendo empregada por usuarios de todo o mundo como un recurso complementario en cursos de doutoramento, en mestrados e mesmo nas súas investigacións. Tendo en conta o uso e a aceptación entre a comunidade internacional, presentamos un paquete de teoría de xogos cooperativos en **R**, `TUGlabR`, unha extensión de `TUGlab` que permite ao usuario traballar con xogos xerais tendo sempre presente a limitación inescusábel da entrada da función característica (vector en $\mathbb{R}^{2^n-1}$, sendo n o número de xogadores). O paso da plataforma xa existente a esta nova ferramenta de **R** supón un avance significativo. Non só se incorporan funcións adicionais, senón que ademais están dispoñibles en aberto, o que facilita o seu acceso a toda a comunidade científica.

Ademais das funcións clásicas de `TUGlab`, `TUGlabR` contén outras funcións relacionadas coas investigacións máis recentes dos autores. Así, por exemplo, é posible calcular o *core-center* dun xogo ([2]), obter os xogos marxinais ou xogos das caras ([8]) ou dar os valores ponderados ([4] e [12]). `TUGlabR` inclúe funcións que permiten analizar particularidades que poden ser engadidas ao xogo en situacións nas que existen prioridades, asimetrías entre xogadores, sistemas coalicionais ou comunicación restrinxida por un grafo.

Por último, utilizaremos `TUGlabR` para calcular a distribución aconsellada polo valor de Shapley na repartición de emisións de CO_2, considerando 40 países europeos. O valor de Shapley obtense, agora, empregando un algoritmo desenvolto en [5].

Palabras e frases chave: xogos coalicionais, comparación de métodos de repartimento, visualización gráfica.

AGRADECEMENTOS

Este traballo está financiado polo proxecto PID2021-124030NB-C33, MCIN/AEI/10.13039/501100011033/, "ERDF A way of making Europe"/EU, e polo "Programa de axudas á etapa predoutoral da Xunta de Galicia", Consellería de Educación, Universidade e Formación Profesional, número ED481A 2021/325.

REFERENCIAS

[1] Cano-Berlanga S., Giménez-Gómez J. M. and Vilella C. (2017). Enjoying cooperative games: The **R** package GameTheory. *Applied Mathematics and Computation* 305, 381-393.

[2] González-Díaz J. and Sánchez-Rodríguez, E. (2007). A natural selection from the core of a TU game: the core-center. *International Journal of Game Theory* 36, 27-46.

[3] Grande Cougil, R.P., Mosquera Rodríguez, M.A. and Ramos Valcarcel, D. (2011). TUGlabWeb: Interfaz Web para xogos TU. *X Congreso Galego de Estatística e Investigación de Operacións, Vigo.*

[4] Kalai, E. and Samet, D. (1987). On weighted Shapley values. *International Journal of Game Theory* 16(3), 205-222.

[5] Le Creurer, I.J., Mirás Calvo, M. A., Núñez Lugilde, I., Quinteiro Sandomingo, C. and Sánchez Rodríguez, E. (2023). On the computation of the Shapley value and the random arrival rule. Avaliable at `https://ssrn.com/abstract=4293746`.

[6] Maschler, M., Zamir, S., and Solan, E. (2020). Game Theory. *Cambridge University Press.*

[7] Mirás Calvo, D. (2008). Programas informáticos orientados a xogos TU. Proyecto fin de máster. Universidade de Vigo.

[8] Mirás Calvo, M. A., Quinteiro Sandomingo, C. and Sánchez Rodríguez, E. (2020). The boundary of the core of a balanced game: face games. *International Journal of Game Theory* 49, 579-599.

[9] Mirás Calvo, M. A. and Sánchez Rodríguez, E. (2006). TUGlab users guide. http://mmiras.webs.uvigo.es//TUGlabGUIDE.pdf

[10] Mirás Calvo, M. A. y Sánchez Rodríguez, E. (2010). Herramientas informáticas de cálculo y representación gráfica para juegos TU. *La Gaceta de la RSME* 13 (1), 89-108.

[11] Núñez Lugilde, I., Mirás Calvo, M. A., Quinteiro Sandomingo, C. and Sánchez Rodríguez, E. (2023). ClaimsProblems: Analysis of Conflicting Claims, **R** package version 0.2.1.

[12] Sánchez Rodríguez, E., Mirás Calvo, M. A., Quinteiro Sandomingo, C. and Núñez Lugilde, I. (2023). Coalitional-weighted Shapley values. *International Journal of Game Theory* (to appear).

[13] Staudacher, J. and Anwander, J. (2019). Using the **R** package CoopGame for the analysis, solution and visualization of cooperative games with transferable utility; R Vignette.

[14] Von Neumann, J. and Morgenstern, O. (2007). Theory of Games and Economic Behavior. *Princeton University Press.*

3D Point Cloud Semantic Segmentation Through Functional Data Analysis

Manuel Oviedo de la Fuente [1], Carlos Cabo [2,3], Javier Roca-Pardiñas [4], E. Louise Loudermilk [5], Celestino Ordóñez [2]

[1]Department of Mathematics, CITIC, Facultad de Informática, University of A Coruña, Campus de Elviña s/n, 15071, A Coruña, Spain

[2] Department of Mining Exploitation and Prospecting, University of Oviedo, Escuela Politécnica de Mieres, 33600, Mieres, Spain

[3] Faculty of Science and Engineering, Swansea University, Singleton Campus, Swansea, UK.

[4] Department of Statistics, The Galician Centre for Mathematical Research and Technology (CITMAga), University of Vigo, Vigo, Spain.

[5] Center for Forest Disturbance Science, USDA Forest Service Southern Research Station, 320 Green Street, 30602 Athens, GA, USA.

ABSTRACT

Here, we propose a method for the semantic segmentation of 3D point clouds based on functional data analysis. For each point of a training set, a number of handcrafted features representing the local geometry around it are calculated at different scales, that is, varying the spatial extension of the local analysis. Calculating the scales at small intervals allows each feature to be accurately approximated using a smooth function and, for the problem of semantic segmentation, to be tackled using functional data analysis. We also present a step-wise method to select the optimal features to include in the model based on the calculation of the distance correlation between each feature and the response variable. The algorithm showed promising results when applied to simulated data. When applied to the semantic segmentation of a point cloud of a forested plot, the results were better than using a standard multiscale semantic segmentation method. The comparison with two popular deep learning models showed that our proposal requires smaller training samples sizes and can compete with these methods in terms of prediction.

Keywords: Laser scanning; Multiscale analysis; Functional data; Multiclass classification; Variable selection.

REFERENCES

Berrendero JR, Cuevas A, Torrecilla JL (2016) Variable selection in functional data classification: a maxima-hunting proposal. Statistica Sinica 26(2): 619-638

Cabo C., Ordóñez C., Sáchez-Lasheras, F., Roca-Pardiñas, J. and de Cos-Juez J (2019) Multiscale Supervised Classification of Point Clouds with Urban and Forest Applications. Sensors 19: 4523.

Oviedo de la Fuente, M., Cabo, C., Roca-Pardiñas, J. et al. 3D Point Cloud Semantic Segmentation Through Functional Data Analysis. JABES (2023). https://doi.org/10.1007/s13253-023-00567-w

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

Xie Y, Tian J, Zhu XX (2020) Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. IEEE Geosci. Remote Sens Mag 8: 38-59

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

COALITION-WEIGHTED SHAPLEY VALUES

Estela Sánchez Rodríguez[1], Miguel Ángel Mirás Calvo[2],
Carmen Quinteiro Sandomingo[3], Iago Núñez Lugilde[1]

[1] CINBIO, Universidade de Vigo, SiDOR. Departamento de Estatística e Investigación Operativa
[2] Universidade de Vigo, RGEAF, Departamento de Matemáticas
[3] Universidade de Vigo, Departamento de Matemáticas

SUMMARY

The (symmetric) Shapley value of a coalitional game, introduced by Lloyd Shapley ([6]), can be viewed as the amount that a player may reasonably expect to get from playing the game. Different characterizations, generalizations, and adaptations of this value have been studied. Shapley himself also considered the possibility of treating players non-symmetrically. The positively weighted Shapley value is defined by introducing exogenous weights to players in order to cover the asymmetries that are not included in the underlying game. The positively weighted Shapley value does not allow zero-weight players. The weighted Shapley value (introduced by [6] and axiomatized by [3]), gets around this drawback by classifying the players into hierarchical classes so that players from a class have relative zero-weight with respect to players that belong to a higher class.

We introduce a new class of values for coalitional games: the coalition-weighted Shapley values. Weights can be assigned to coalitions, not just to players, and zero-weights are admissible. The Shapley value belongs to this class. Coalition-weighted Shapley values recommend for each game the allocation defined by the Shapley value of a weighted game obtained as a linear convex combination of the associated marginal games (or face games, see [2] and [4]). Coalition-weighted Shapley values are random order values, Harsanyi values, and multiweighted Shapley values ([1]). Positively weighted Shapley values and weighted Shapley values can be seen as the limit of a sequence of iterated coalition-weighted Shapley values. We provide axiomatic characterizations of coalition-weighted Shapley values through properties that do not involve the weights. Finally, we discuss how to extend our model to include exogenous coalition structures as in the hierarchical and Owen values.

Keywords: Shapley value, weighted Shapley values, marginal games, characterization of values, coalition structure

ACKNOWLEDGEMENTS

This work was supported by project PID2021-124030NB-C33 that is funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"/EU, and by grant ED481A 2021/325 funded by *programa de axudas á etapa predoutoral da Xunta de Galicia, Consellería de Educación, Universidade e Formación Profesional*.

REFERENCES

[1] Besner, M. (2020) Parallel axiomatizations of weighted and multiweighted Shapley values, random order values, and the Harsanyi set. *Social Choice and Welfare*, 55, 193–212.

[2] González Díaz J. and Sánchez Rodríguez, E. (2008) Cores of convex and strictly convex games. *Games and Economic Behavior*, 62(1), 100–105.

[3] Kalai, E. and Samet, D. (1987) On weighted Shapley values. *International Journal of Game Theory*, 16, 205–222.

[4] Mirás Calvo, M. A., Quinteiro Sandomingo, C., and Sánchez Rodríguez, E. (2020) The boundary of the core of a balanced game: face games. *International Journal of Game Theory*, 49, 579–599.

[5] Sánchez Rodríguez, E., Mirás Calvo, M. A., Quinteiro Sandomingo, C., and Núñez Lugilde, I. (2023) Coalition-weighted Shapley values. *International Journal of Game Theory*. To appear.

[6] Shapley, L. S. (1953) A value for n-person games. *in Contributions to the Theory of Games*, ed. by H. Kuhn and A. Tucker, Princeton: Princeton University Press, Annals of Mathematical Studies, 307–317.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

DECONSTRUCTING THE CORE OF THE AIRPORT GAME WITH SYMMETRIC PLAYERS

Carmen Quinteiro Sandomingo[1], Miguel Ángel Mirás Calvo[2],
Estela Sánchez Rodríguez[3]

[1] Universidade de Vigo, Departamento de Matemáticas
[2] Universidade de Vigo, RGEAF, Departamento de Matemáticas
[3] CINBIO, Universidade de Vigo, SiDOR. Departamento de Estatística e Investigación Operativa

SUMMARY

The airport problem, introduced by [2], is a classic cost allocation problem. It is based on modeling situations in which a set of agents are linearly ordered by their needs in such a way that satisfying the need of one agent implies satisfying all those with lower needs. Although the initial application was with airlines needing different runway sizes, its scope is really much wider. A complete survey on airport problems is [3].

In this paper we study the impact of symmetric players (players with the same need) in the core of the airport game. In fact, the number of symmetric players determines the number of redundant constraints in the core, which makes the core simpler and more regular. Exact expressions for the Shapley value and the nucleolus for this kind of problems are given in [4]. We show that when there are n agents distributed in t groups of symmetric agents, the core-center ([1]) of the $n-1$ dimensional core can be computed through the mixed moments of the $t-1$ dimensional uniform variables where only one representative from each group is needed. We also study the convergence of some rules under replication of an airport problem.

Keywords: cooperative TU games, core, core-center, airport games, symmetric players, mixed moments, replication

ACKNOWLEDGEMENTS

This work was supported by project PID2021-124030NB-C33 that is funded by MCIN/AEI/10.13039/501100011033/, and by "ERDF A way of making Europe"/EU.

REFERENCES

[1] González-Díaz, J., Mirás Calvo, M. A., Quinteiro Sandomingo, C., and Sánchez Rodríguez, E. (2016). Airport games: The core and its center. *Mathematical Social Sciences* 82, 105–115.

[2] Littlechild, S. C., and G. Owen (1973). A simple expression for the Shapley value in a special case. *Management Science* 20, 370–372.

[3] Thomson, W. (2014). Cost allocation and airport problems. Rochester Center for Economic Research Working Paper.

[4] Owen, G. (1995). Game Theory. Academic Press.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Kernel estimators of the ROC curve for functional biomarkers

Graciela Estévez-Pérez[1]

[1]Departamento de Matemáticas, Universidade da Coruña

ABSTRACT

Estévez and Vieu (2021) developed recently a diagnostic test that use functional variables as biomarkers and proposed an empirical estimate of the functional ROC curve. In order to improve this methodology, the present work introduces a procedure to obtain a smooth version of nonparametric estimator of ROC curve. In addition, a simulation study is carry out to investigate the discriminatory and predictive abilities of the new functional diagnostic tests, paying special attention to the improvement achieved using the kernel estimators versus the empirical one. Finally, to illustrate the proposed methodology, the analysis of a real medical data set is performed.

Keywords: Diagnostic test; Functional biomarker; Kernel smoothing; ROC curve; Preorder relation.

1. INTRODUCTION

Technical development over the last few decades has resulted in the emergence of complex data, in many cases functional data (FD). This type of data can emerge in many medical studies which are geared towards detecting diseases, predicting their course or evaluating the response to a therapy, to name a few. Thus, it is very useful to have statistical methods enabling us to evaluate diagnostic tests based on functional biomarkers.

The receiver operating characteristic (ROC) curve is the most widely used tool for evaluating diagnostic tests that consist in a continuous measurement (Pepe, 2004). Certainly it is very versatile. As Gonçalves et al. (2014) point out, *the ROC analysis is useful to: (i) evaluate the discriminatory ability of a continuous marker to correctly assign into a two-group classification; (ii) find an optimal cut-off point to least misclassify the two-group subjects; and (iii) compare the efficacy of two (or more) diagnostic tests or markers*. The ROC curve plots the sensitivity or true positive rate (TPR) versus one minus the specificity or false positive rate (FPR) across all the possible decision thresholds or cutoffs. More details about ROC analysis can be found in the books by Pepe (2004) and Zhou et al. (2009).

Any diagnostic procedure based on ROC curves needs to face the question of estimating the distribution function of the considered random biomarker. Therefore, the extension of existing multivariate approaches to functional setting is far from being direct. Estévez and Vieu (2021) developed recently a ROC curves method for analysing functional data themselves. The authors investigated a diagnostic test that use functional variables as biomarkers and proposed a functional version of ROC analysis, which is linked to a suitable way to rank the sample of functional data. This methodology led to an empirical estimate of the functional ROC curve, which presents the obvious weakness of being a step function. In this work, a procedure to obtain a smooth version of nonparametric estimator of ROC curve introduced by Estévez and Vieu (2021) is proposed. Firstly, the new functional diagnostic test is presented by introducing several kernel approaches of the functional ROC curve estimator. Then, it is conducted a simulation study to assess the finite-sample properties of the proposed estimators and it is illustrated the use of the introduced functional diagnostic test through an example with real medical data.

2. FUNCTIONAL DIAGNOSTIC TEST

Let χ be a functional biomarker and χ_1, χ_2 be the conditioned functional variables by affected subjects (A) and non-affected ones (NA), which are valued in a semi-metric space (E, d).

We are in front of a discrimination problem with n_1 and n_2 i.i.d. random samples from χ_1 and χ_2, respectively

$$\{\chi_{ij}(t); i = 1, 2; j = 1, ..., n_i; t \in T\}$$

where, without loss of generality, $T = [0, 1]$ and $n = n_1 + n_2$ denotes the total sample size.

The diagnostic method defined in Estévez and Vieu (2021) consists in finding a cutoff curve or discrimination curve $\chi_{c_0} \equiv \chi_{c_0}(t) \in E$, with $t \in T$, which separates affected (A) and non-affected (NA) subjects as much as possible. The procedure can be split into 4 steps:

1. **Construct the totally ordered set** $(E_1, \preceq_1)$:

$$E_1 = \{\chi_c(t) = cM_A(t) + (1 - c)M_{NA}(t); t \in T, c \in \mathbb{R}\} \tag{1}$$

   where $M_A \equiv M_A(\cdot)$ and $M_{NA} \equiv M_{NA}(\cdot)$, curves of the functional space E, are two curves representatives of each group A and NA, respectively, and $\preceq_1$ is the total order in E_1 induced by the natural ordering in $\mathbb{R}$.

2. **Establish a preorder relation in** E **(**$\preceq$**).** From a projection function $\eta : E \to E_1$, as for example,

$$\eta(x) = \min\{y \in E_1 / d(x, y) = \min\{d(x, y_1); y_1 \in E_1\}\},$$

   that assigns each curve x in E the nearest element in E_1, two arbitrary functional elements x_1 and x_2 in E can be preordered: x_1 precedes to x_2 in E (which is written as $x_1 \preceq x_2$), if and only if $\eta(x_1) \preceq_1 \eta(x_2)$ in E_1.

3. **Define the functional ROC curve (FROC)** by means of representation:

$$\{(\alpha_c, 1 - \beta_c), c \in \mathbb{R}\} = \{(1 - F_{\eta_2}(\chi_c), 1 - F_{\eta_1}(\chi_c)), c \in \mathbb{R}\} \tag{2}$$

   with $F_{\eta_i}(x) = P(\eta(\chi_i) \preceq_1 x); x \in E_1$, the distribution functional of $\eta(\chi_i)$, $i = 1, 2$.

4. **Select as optimal cutoff curve** the functional element χ_{c_0}, such that the pair $(1-\alpha_{c_0}, 1-\beta_{c_0})$ is as close as possible to $(1, 1)$ (*North-West corner criterion*).

In this paper it is proposed to make a slight amendment in this method, specifically in the third step. If X_1^c and X_2^c denote the real-valued random variables of the projection coefficients in E_1 for groups A and NA, respectively, and F_1, F_2 their corresponding distribution functions, we have that, $\forall x \in E_1$

$$F_{\eta_i}(x) = F_{\eta_i}(\chi_c) = P(\eta(\chi_i) \preceq_1 \chi_c) = P(X_i^c \leq c) = F_i(c); i = 1, 2$$

because all functional element $x \in E_1$ can be written as $x = \chi_c = cM_A + (1 - c)M_{NA}$, for some $c \in [0, 1]$.

Therefore, the FROC curve (2) can be written as

$$\{(\alpha_c, 1 - \beta_c), c \in \mathbb{R}\} = \{(1 - F_2(c), 1 - F_1(c)), c \in \mathbb{R}\} \tag{3}$$

Note that, from functional biomarker χ, applying the projection function $\eta : E \to E_1$, a real biomarker X^c to discriminate affected (A) and non-affected (NA) subjects is obtained. As the **FROC curve of** χ **matches the ROC curve of real variable** X^c**, being** X^c **the value** $c \in \mathbb{R}$ **such that** $\eta(\chi) = cM_A + (1 - c)M_{NA}$, the estimation of functional ROC curve (FROC) can be derived from the estimation of a real ROC curve.

3. NONPARAMETRIC ESTIMATION OF FUNCTIONAL ROC CURVE

The simplest nonparametric method for estimating the functional ROC curve is the empirical estimator (Estévez and Vieu, 2021), which is based on plugging in empirical distributions of affected and non-affected into (3). The empirical version of FROC (**ROC.EMP**) is given by

$$\{(\widetilde{\alpha}_c, 1-\widetilde{\beta}_c), c \in \mathbb{R}\} = \{(1-\widetilde{F}_2(c), 1-\widetilde{F}_1(c)), c \in \mathbb{R}\},$$

where $\widetilde{F}_i(c) = \frac{1}{n_i}\sum_{j=1}^{n_i} I[X_i^c \leq c]; i = 1, 2$ is the empirical distribution function of F_i $(i = 1, 2)$.

As with the empirical estimator, the simplest way of getting a kernel estimation of functional ROC curve is to consider some kernel smoother of the ROC curve for the real variable X^c. A review of the literature on non-parametric estimation of ROC curves of real biomarkers has been led, getting some of the methods mostly employed in practice:

ROC.K.Ll98 (Lloyd, 1998) It uses kernel estimates directly for F_1 and F_2 and it has better mean square error (MSE) than the ROC.EMP.

ROC.K.ZH02 It is based on the kernel smoothing method originally proposed by Altman (1995) for estimation of the distribution function.

ROC.K.HH03 (Hall and Hyndman, 2003). It estimates the ROC curve as a function and provides a substantial improvement in MISE over other proposed methods, especially when the two distributions are very different.

ROC.K.P16 (Pulit, 2016) It is based on the idea of estimating the ROC curve as a distribution function. This estimator is invariant under nondecreasing data transformations and has better asymptotic MSE properties than other estimators involving kernel smoothing.

Therefore, it is proposed the following kernel estimator for the functional ROC curve:

$$\{(\widehat{\alpha}_{c,h}, 1-\widehat{\beta}_{c,h}), c \in \mathbb{R}\} = \{\widehat{ROC}_h(c), c \in \mathbb{R}\},$$

where $\widehat{ROC}_h$ is one of the four kernel estimators of the ROC curve for the real variable X^c described above (**ROC.K.Ll98**, **ROC.K.ZH02**, **ROC.K.HH03** or **ROC.K.P16**) and h is the uni- or two-dimensional vector of bandwidth parameters according to the choice.

4. SIMULATION STUDY

The empirical study was undertaken to examine the impact of estimation method of ROC curve in the functional context and their interaction with the other parameters involved in diagnostic test: type of representative curves ($r.t$), choice of ordered set (E_1) and semi-metric for the projection ($d.t$). This effect is measured by on the discriminatory power of the test, through numerical indexes like the AUC, and on its predictive ability through misclassification rates. This study is a continuation of the Estévez and Vieu (2021) one, then, the interested reader into the general conditions of the study (scenarios and design parameters) is referred to this paper.

According to the diagnostic test outlined in Section 2, to build functional ROC curves we need first and foremost to project the sample curves on the ordered subset $(E_1, \preceq_1)$. For that, we consider E_1 defined as in (1) and as values of $r.t$ (type of representative curve), $r.t = 1$ (functional mean) and, among the depthmeasures, $r.t = 3$ (depth.mode) because their values of AUC, specificity and sensitivity presented lower variability (see Estévez and Vieu, 2021). On $d.t$ parameter (type of semi-metric), we have selected $d.t = 1$ (L2), $d.t = 2$ (Semi.Basis), $d.t = 3$ (CCC), $d.t = 7$ (Semi.Fourier) and $d.t = 8$ (Semi.mplsr). Finally, for each combination of representative curve and

semi-metric, the non-parametric estimation of functional ROC curve will be built for every one of five estimators outlined in Section 3. Where necessary, the Epanechnikov kernel will be used and the bandwidth parameters will be selected as pointed in the aforementioned section. For more details, see their references.

Part of a comprehensive simulation study, which was carried out to investigate the discriminatory and predictive abilities of the new functional diagnostic tests, is shown below. In particular, we have kept to the scenario F1. Table 1 and Figures 1 and 2, and others analogous for the several combinations of parameters, were built and analysed. A first general conclusion, already noted in Estévez and Vieu (2021), is that FMean curve ($r.t = 1$) yields more effective results than the depthmeasure depth.mode ($r.t = 3$), though in some cases the results only differ slightly (Table 1). With regard to the semi-metrics, the Semi.Fourier ($d.t = 7$) provides the worst results for each estimation method considered and the L2 ($d.t = 1$) shows a competitive performance when variability is not high. The best outcomes come from the semi-metrics CCD ($d.t = 3$), Semi.Basis ($d.t = 2$) and Semi.mplsr ($d.t = 8$), being the first one which yields the most successful results.

Table 1: Means and standard deviations (in parenthesis) of area under ROC curve (AUC) for several values of $r.t$ and $d.t$ when F1 scenario is considered. Sample size $n_1 = n_2 = 20$ and $\sigma = \sigma_1 = 0.06$

	$d.t = 1$	$d.t = 2$	$d.t = 3$	$d.t = 7$	$d.t = 8$	**Method**
$r.t = 1$	0.9419 (0.0308)	0.9482 (0.0295)	**0.9594 (0.0256)**	0.8855 (0.0442)	0.9518 (0.0296)	**ROC.EMP**
$r.t = 3$	0.9391 (0.0337)	0.8833 (0.0726)	0.9454 (0.0370)	0.7877 (0.1302)	0.9477 (0.0297)	
$r.t = 1$	0.9246 (0.0336)	0.9353 (0.0314)	**0.9497 (0.0282)**	0.8612 (0.0423)	0.9366 (0.0339)	**ROC.K.Ll98**
$r.t = 3$	0.9243 (0.0350)	0.8638 (0.0722)	0.9339 (0.0375)	0.7612 (0.1274)	0.9343 (0.0354)	
$r.t = 1$	0.9152 (0.0353)	0.9275 (0.0327)	**0.9429 (0.0298)**	0.8514 (0.0432)	0.9270 (0.0376)	**ROC.K.ZH02**
$r.t = 3$	0.9154 (0.0367)	0.8549 (0.0719)	0.9264 (0.0387)	0.7536 (0.1234)	0.9266 (0.0370)	
$r.t = 1$	0.9345 (0.0313)	0.9450 (0.0289)	**0.9580 (0.0261)**	0.8724 (0.0424)	0.9463 (0.0319)	**ROC.K.HH03**
$r.t = 3$	0.9340 (0.0331)	0.8745 (0.0733)	0.9426 (0.0360)	0.7679 (0.1307)	0.9438 (0.0340)	
$r.t = 1$	0.9415 (0.0220)	0.9480 (0.0180)	**0.9553 (0.0163)**	0.8976 (0.0335)	0.9486 (0.0209)	**ROC.K.P16**
$r.t = 3$	0.9408 (0.0230)	0.8963 (0.0609)	0.9454 (0.0239)	0.8043 (0.1245)	0.9463 (0.0214)	

For the sake of comparison, the estimator ROC.K.P16 has proved to be the best option (higher areas with the least amount of variability) except for very low values of σ and σ_1. In these cases (see Figure 2), all other kernel methods are more competitive. The results display a hierarchy, from more to less efficacy, between the other kernel methods: ROC.K.HH03 is the first one, then is ROC.K.Ll98, and finally is ROC.K.ZH02. In addition, ROC.K.HH03 is the closest rival of ROC.EMP and goes even further in when sample size is high and σ and σ_1 are low.

Since the analysis of optimal coefficients of sensitivity and specificity led to similar conclusions, from the discriminatory capability point of view, the best election for F1 scenario is $r.t = 1$ as representative curve, $d.t = 3$ as semi-metric and ROC.K.P16 as kernel estimator of functional ROC curve. When there is more overlap in the functional data, i.e., affected and non-affected groups are more difficult to distinguish, ROC.K.HH03 will be a good option.

From the classification ability point of view, we carried out the same trial that Estévez and Vieu

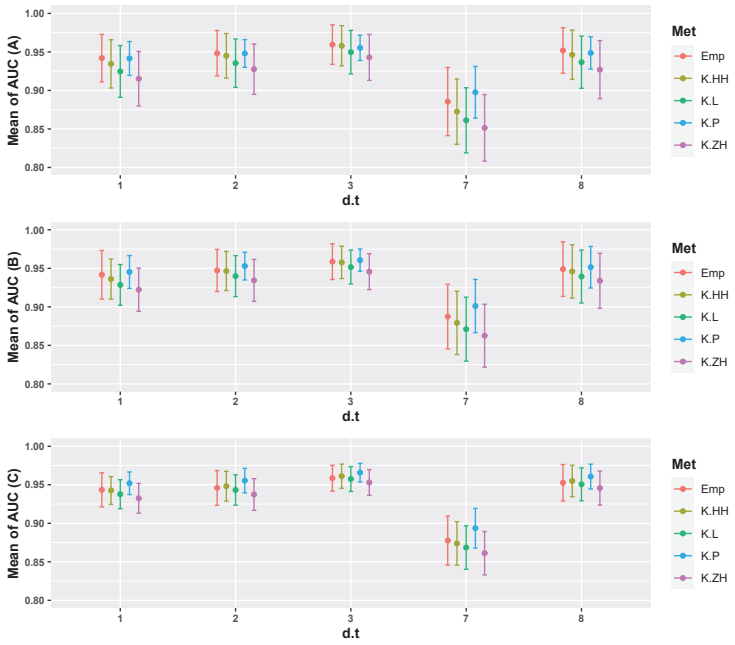


Figure 1: Scenario F1. Means of area under ROC curve (AUC) for several values of $d.t$ when $r.t = 1$ is considered. Values $\sigma = 0.06$ and $\sigma_1 = 0.06$, and sample size $n_1 = n_2 = 20$ ((A) top chart); $n_1 = n_2 = 30$ ((B) central chart); $n_1 = n_2 = 50$ ((C) bottom chart)

(2021) for the combination of parameters under consideration. For each replication and each combination of parameters (n_1, n_2) and (σ, σ_1), two samples were considered: the training sample and the test sample whose origin group is known. For these choices, the means and standard deviations of the misclassification rates, obtained by classifying 10 test samples (100 replications), were calculated.

Table 2 reports the mean values considering $r.t = 1$ and those values of $d.t$ with the best performance to discriminate between affected and non-affected ($d.t = 2, 3$ and 8). The results bring to light the increased misclassification percentages when σ and σ_1 increase and when n_1 and n_2 decrease. Regarding the estimation method, the semi-metrics resulting in improved performance are $d.t = 3$ or $d.t = 8$ for ROC.EMP estimator; $d.t = 3$ for kernel estimators ROC.K.Ll98, ROC.K.ZH02 and ROC.K.HH03 in all cases, except when the values of variability are low ($\sigma = \sigma_1 = 0.04$) in which case $d.t = 2$ presents a marginal improvement particularly for small samples; $d.t = 3$ is also the best option for ROC.K.P16 although the results are less conclusive. Finally, we note that the kernel estimators ROC.K.Ll98, ROC.K.ZH02 and ROC.K.HH03 provide the lowest misclassification rates, very close to each other. They are substantially better than the empirical method (ROC.EMP) and the Pulit's one (ROC.K.P16). This latter procedure only provides competitive rates for the highest values of variability, although the previous results have proved that it is the best estimator to discriminate between groups (Table 1 and Figures 1 and 2).

5. REAL DATA ANALYSIS

To illustrate the methodology of functional diagnostic test and to complete the comparison among kernel estimation methods (ROC.K) and the empirical ROC curve (ROC.EMP), the analysis of a real data set is presented in this section. The diffusion tensor imaging scan dataset (DTI) includes

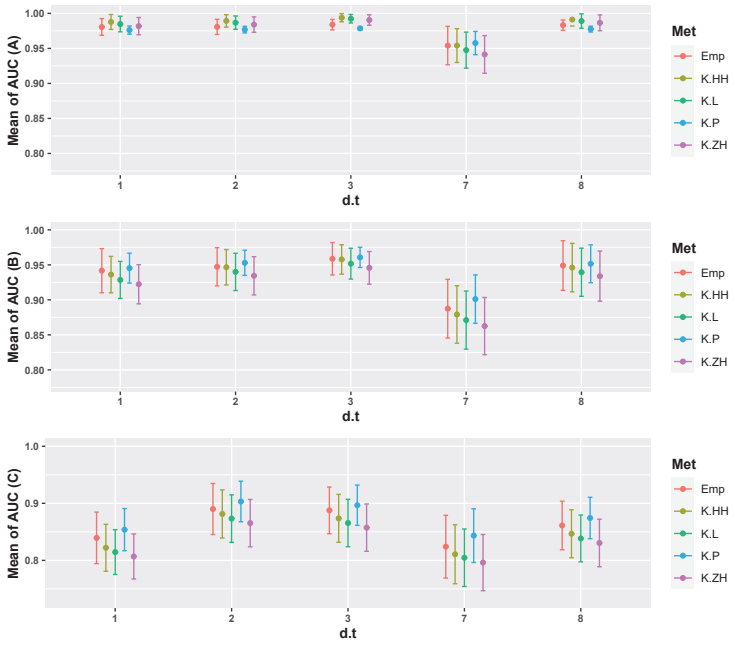


Figure 2: Scenario F1. Means of area under ROC curve (AUC) for several values of $d.t$ when $r.t = 1$ is considered. Sample size $n_1 = n_2 = 30$, and $\sigma = 0.04$ and $\sigma_1 = 0.4$ ((A) top chart); $\sigma = 0.06$ and $\sigma_1 = 0.06$ ((B) central chart); $\sigma = 0.08$ and $\sigma_1 = 0.1$ ((C) bottom chart)

cerebral white matter tracts for multiple sclerosis (MS) patients and healthy controls. MS is a disease of the central nervous system that causes lesions in white matter tracts, thus disrupting the transmission of signals and resulting in severe patient disability. DTI is a magnetic resonance imaging modality that measures diffusion of water in the brain, allowing the quantification of white matter tract integrity. From DTI scans, fractional anisotropy (FA) measurements are extracted which provides information about water diffusion in the brain (values near one indicate complete anisotropy and values near zero indicate completely isotropic diffusion), and thus about MS diagnosis and progression. That is, FA is used as a marker of disease progression in MS.

The DTI dataset, which can be found in the R-package `refund` (Goldsmith et al., 2016), contains FA measurements of healthy and diseased individuals, recorded at several locations along the callosal fibre tract and the right corticospinal tract in the brain. Because we observe tract profiles for each subject over time, the DTI dataset can be considered a functional dataset. The whole sample was randomly divided into a training sample (60 MS and 30 controls) and a test sample (24 MS and 12 controls) with the aim of gauging the predictive quality of the diagnostic test. In panel (A) of Figure 3 the observed FA measurements for training sample are shown, differentiating individuals affected and non-affected with MS. Note that, in general, MS patients present lower FA measurements than healthy individuals.

The functional diagnostic method for DTI dataset has been applied and the Figure 3 shows the result. The representative curve of group has been the functional mean ($r.t = 1$) and as semi-metric we have chosen $d.t = 1$ (L.2) because it provided the best results of discrimination. For these parameters, the ordered subset E_1 and the projections over it are shown in panels (B), (C) and (D) of Figure 3. The corresponding estimators of the FROC curve for the methods, which have delivered the best results (ROC.EMP, ROC.K.HH03 and ROC.K.P16), are presented in panel

Table 2: Scenario F1. Means of the misclassification percentages considering $r.t = 1$ and those values of $d.t$ with the best performance to discriminate between affected and non-affected.

	$d.t = 2$			$d.t = 3$			$d.t = 8$			
(σ, σ_1)	20	30	50	20	30	50	20	30	50	**Method**
$(0.04, 0.04)$	19.15	15.2	10.4	19.95	12.45	9.15	14.1	12.4	12	
$(0.06, 0.06)$	18.1	14.65	13.75	17.9	12.95	12.4	17.7	12.35	13.3	**ROC.EMP**
$(0.08, 0.1)$	24.1	23.1	21.3	24.75	21.05	20.65	26.8	23.65	22.65	
$(0.04, 0.04)$	**7.25**	**6.65**	**6.45**	8.6	7.25	6.6	11.25	10.2	8.95	
$(0.06, 0.06)$	12.0	12.5	12.25	**11.2**	**10.05**	**8.95**	12.7	10.8	9.8	**ROC.K.Ll98**
$(0.08, 0.1)$	**20.75**	21.8	20.35	22.85	**20.45**	**19.25**	24.9	22.05	21.55	
$(0.04, 0.04)$	**7.1**	**6.55**	**6.25**	8.35	7.4	6.85	11.65	10.1	8.95	
$(0.06, 0.06)$	12.4	12.4	12.1	**11.15**	**10.25**	**9.1**	12.7	10.65	9.65	**ROC.K.ZH02**
$(0.08, 0.1)$	**20.8**	22.0	20.25	22.85	**20.55**	**19.25**	25.05	22.15	21.45	
$(0.04, 0.04)$	**6.85**	**6.7**	**6.4**	8.35	7.4	6.65	11.15	10.1	8.85	
$(0.06, 0.06)$	12.05	12.35	12.15	**10.9**	**9.85**	**9.2**	12.65	10.85	9.95	**ROC.K.HH03**
$(0.08, 0.1)$	**21.1**	21.75	20.6	22.95	**20.2**	**19.45**	25.1	21.9	21.6	
$(0.04, 0.04)$	18.1	12.2	8.45	18.6	10.85	8.4	16.5	12.0	9.45	
$(0.06, 0.06)$	16.35	13.6	13.0	17.15	11.55	9.85	16.55	12.55	10.6	**ROC.K.P16**
$(0.08, 0.1)$	22.7	22.45	20.2	23.5	**20.4**	**19.55**	25.85	22.15	21.85	

(F) of Figure 3, and the optimal discrimination curve for ROC.K.P16 together with the projected curves in E_1 can be seen in panel (G). The AUC values were $0.878, 0.855, 0.887$, respectively.

Finally, in order to evaluate the classification ability, we consider the test samples, with a known origin group, both for MS patients and for controls. For methods ROC.EMP, ROC.K.HH03 and ROC.K.P16 the reached misclassification percentages were respectively $29.2\%, 37.5\%, 37.5\%$ in affected and $16.7\%, 16.7\%, 16.7\%$ in non-affected.

6. SOME CONCLUSIONS

The following key conclusions can be drawn:

- The representative curve FMean ($r.t = 1$) appears to be the best option. Maybe in presence of outliers, some more robust measure might be preferable, for example depth.mode ($r.t = 3$), which also provides good results as shown by the simulations outcomes.

- As usual in the functional context, the choice of semi-metric plays a major role and it depends mostly on the shape of data. In this case, $d.t = 2$ or $d.t = 3$ were the optimal options in general. The semi-metric mplsr ($d.t = 8$) has not been optimal in any scenario but it achieves good results in every case. In real data analysis, we chose the one which provides the best results of discrimination in terms of AUC.

- The kernel methods for functional ROC curve estimation have a major advantage of yielding a smooth ROC curve, unlike an empirical ROC curve. Additionally, they are more competitive than the empirical one, especially the methods K.P16 and K.HH03. K.P16 represents the best option from discrimination point of view but K.HH03 shows also the best outcomes in terms of misclassification rates.

- When variability is low, the empirical method underestimates the optimal cutoff point (c_0) obtained by means of the *North-West corner criterion*, yielding high misclassification percentages in the non-affected group. Maybe the choice of a different criteria to select the optimal cutoff point, as the one based on Youden index, could help improve this aspect.

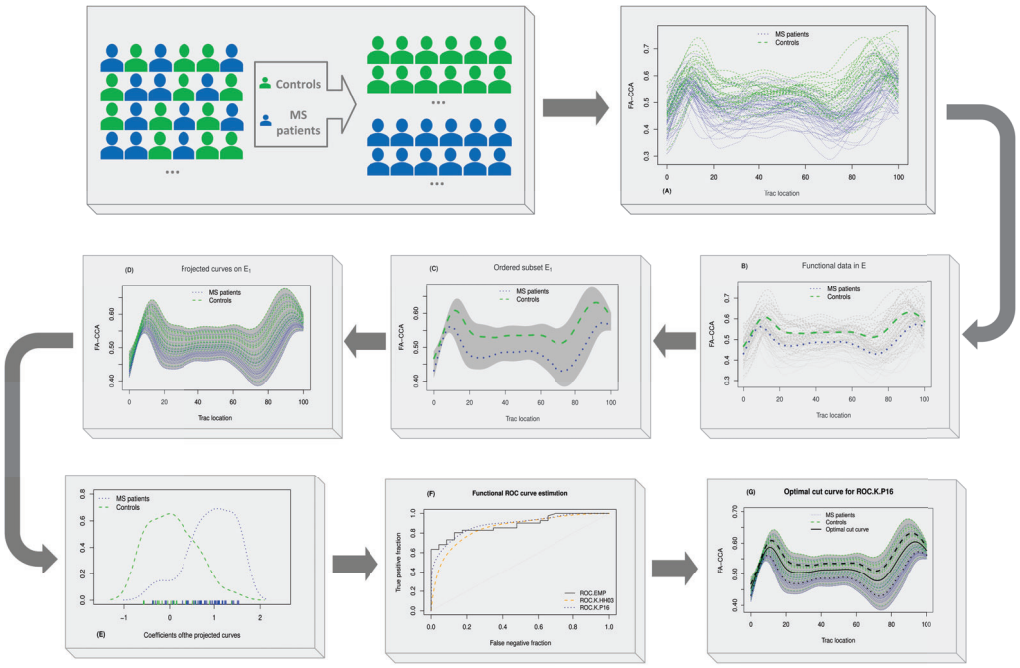


Figure 3: Process of functional diagnostic method for `DTI` dataset. Smoothed curves for affected individuals (blue) and healthy individuals (green) (A). Representative curves $r.t = 1$ by group (B); Ordered subset E_1 obtained from functional means (C); Projected curves over E_1 for affected individuals (blue) and healthy individuals (green) (D); Distributions of coefficients for the projected curves in E_1 (E); Estimations of FROC curve for tree methods: ROC.EMP, ROC.K.HH03 and ROC.K.P16 (F); Optimal discrimination curve in E_1 for ROC.K.P16 (G).

The previous remarks open the door to new research linked to the parameters effect on the discriminatory capability in complex situations, for example, in presence of outliers, for clearly heteroscedastic data or unbalanced sample sizes. These challenges will be the target of future research.

FUNDING

This research was supported by MICINN, Spain grant PID2020-113578RB-I00 and by the 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.

REFERENCES

Altman, N. and Leger, C. (1995) Bandwidth selection for kernel distribution function estimation. Journal of Statistical Planning and Inference, 46(2), 195–214.

Estévez-Pérez, G. and Vieu, P. (2021) A new way for ranking functional data with applications in diagnostic test. Comput Stat, 36(1), 127–154.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., ... and Reiss, P. T. (2016) Refund: Regression with functional data. *R package version 0.1-16, 572*.

Gonçalves, L., Subtil, A., Oliveira, M. R. and de Zea Bermudez, P. (2014) ROC curve estimation: An overview. REVSTAT-Statistical journal, 12(1), 1–20.

Hall, P. G. and Hyndman, R. J. (2003) Improved methods for bandwidth selection when estimating ROC curves. Stat Probab Lett, 64(2), 181–189.

Lloyd, C. J. (1998) Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. J Am Stat Assoc, 93(444), 1356–1364.

Pepe, M. S. (2003) The statistical evaluation of medical tests for classification and prediction. Oxford university press.

Pulit, M. (2016) A new method of kernel-smoothing estimation of the ROC curve. Metrika, 79(5), 603–634.

Zhou, X. H., McClish, D. K. and Obuchowski, N. A. (2009) Statistical methods in diagnostic medicine. John Wiley & Sons.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

Regresión Modal con Datos Faltantes

Tomás R. Cotos Yañez [2], Rosa M. Crujeiras [1] e Ana Pérez González [2]

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

[2]Centro de Investigación e Tecnoloxía Matemática de Galicia, CITMAga. Dpto. Estatística e Investigación Operativa Universidade de Vigo

Palabras e frases chave: Algoritmo Mean-Shift, datos faltantes, imputación, regresión modal.

RESUMO

A regresión Modal é un método de estimación de modas locais da distribución condicional dunha variable resposta Y condicionada a $X = x$. É un método alternativo á clásica regresión en media que cobrou moita importancia nas últimas décadas pola súa idoneidade cando a distribución condicional ten colas pesadas, non é simétrica ou presenta máis dunha moda. Tamén polas distintas aplicacións que se poden derivar do seu estudo, por exemplo obter a estructura de agrupación dos datos cando hai múltiples curvas modais. Por outra banda, con frecuencia atopámonos con mostras incompletas, o estudo do comportamento de calquera estimador baixo esta ausencia é crucial para tomar decisións sobre as distintas metodoloxías de datos faltantes que se poden aplicar. O noso traballo ten como obxectivo analizar o comportamento da regresión modal cando hai ausencia de resposta nalgunhas observacións.

Matemáticamente, a regresión modal obtense da maximización da distribución condicional $Y|X = x$. Imos a considerar un escenario multimodal, e polo tanto definimos o conxunto modal condicional como

$$M(x) = \left\{ y : \frac{\partial}{\partial y} f(y|x) = 0, \frac{\partial^2}{\partial^2 y} f(y|x) < 0 \right\}$$

donde $f(y|X = x)$ representa a densidade condicional.

Para estimar de forma non paramétrica o conxunto $\hat{M}_n(x)$, podemos utilizar un estimador kernel da función de densidade $\hat{f}$ e un algoritmo iterativo coma o de *Mean-Shift* proposto por Einbeck and Tutz (2006).

$$\hat{M}(x) = \left\{ y : \frac{\partial}{\partial y} \hat{f}(y|x) = 0, \frac{\partial^2}{\partial^2 y} \hat{f}(y|x) < 0 \right\}$$

No traballo de Chen e outros (2016) estúdase con detalle este algoritmo así como propiedades teóricas do estimador proposto $\hat{M}_n(x)$.

O problema que nos ocupa é cando a resposta Y pode ter observacións faltantes pero a covariable X é totalmente observada. Para modelar esta perda matemáticamente introducimos unha nova variable indicadora, δ, que toma os valores $\{1, 0\}$ nos casos de observar Y ou ser un dato faltante respectivamente. Seguindo a terminoloxía clásica de datos faltantes, supoñemos que o modelo de datos faltantes é *Missing at Random* (MAR) (Rubin, 1976), de forma que:

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = p(X) \tag{1}$$

A nosa proposta consiste en aplicar distintas metodoloxías de datos faltantes e adaptalas ao *Mean-Shift* algoritmo citado anteriormente. As distintas estimacións analizadas neste traballo son:

- *Estimación simplificada*: utiliza o algoritmo só coa parte completa da mostra.
- *Estimación inversamente ponderada*: pondera cada observación pola inversa da probabilidade de dato faltante (1) para estimar a función de densidade.
- *Estimación sobre a mostra imputada*: realízase unha imputación condicional das observacións de Y e a continuación aplícase o algoritmo *Mean-Shift* á mostra completada.
- *Estimación baseada en técnicas de imputación múltiple*: mediante técnicas de imputación baseadas na densidade condicional, impútanse varias veces cada observación faltante. Ao contrario que no caso da media, aquí para obter una estimación final non é recomendable promediar as estimacións derivadas de eses conxuntos completos. Proponse un novo mecanismo baseado no concepto da moda para obter unha estimación final.

Un dos puntos fundamentais da estimación non paramétrica do conxunto $M(x)$ é a selección do parámetro ancho de banda para a estimación tipo kernel da función de densidade, $\hat{f}$. No traballo de Zhou e Huang (2019) realízase unha comparación de diversos estimadores de este parámetro dende un punto de vista computacional. O método de validación cruzada adaptado á estimación de $M(x)$ parece ser unha boa opción.

Mediante un amplo estudo de simulación estudamos o comportamento das técnicas citadas anteriormente. O estudo faise baixo diferentes modelos de datos faltantes e distintas distribucións que proporcionan un abano de modelos unimodais e bimodais.

Agradecementos

Este traballo foi realizado gracias ao finanzamento económico dos seguintes proxectos; PID2020-116587GB-I00(CoDyNP) e PID2020-118101GB-I00 do Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033). Tamén queremos agradecer as axudas a grupos de Referencia competitivos ED431C 2021/24 e do grupo SIDOR (ED431C 2016/040) financiados pola Consellería de Cultura, Educación e Universidade.

REFERENCIAS

Chen Y., Genovese C.R., Tibshirani R.J., Wasserman, L. (2016) Nonparametric modal regression. Ann Stat 44, 489–514.

Einbeck J., Tutz G. (2006) Modelling beyond regression functions: an application of multimodal regression to speed–flow data. J R Stat Soc Ser C-Appl Stat 55 (4):461–475.

Rubin, D. (1976) Inference and Missing Data. Biometrika 63 (3): 581–592.

Zhou H., Huang X. (2019) Bandwidth selection for nonparametric modal regression. Commun Stat-Simul Comput 48 (4) 968–984.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

ANÁLISIS ESTADÍSTICO DE LA VARIACIÓN DE LA DURACIÓN DEL DÍA

Álvarez Hernández, M.[1], Junguito Marcos, I.[2] y Folgueira López, M.[3]

[1] Centro Universitario de la Defensa en la Escuela Naval Militar, área de Matemáticas.
[2] Armada Española, Ministerio de Defensa.
[3] Universidad Complutense de Madrid, departamento de Física de la Tierra y Astrofísica

RESUMEN

El estudio de la rotación de la Tierra es una de las cuestiones de mayor interés tanto en Astronomía como en Geodesia. De manera particular, la perturbación de la rotación terrestre influye directamente en la duración del día solar, afectando, en consecuencia, a la determinación del tiempo universal. Dada la importancia que tiene en la navegación el conocer de forma precisa esta variable, el presente trabajo tiene como propósito realizar un análisis desde el punto de vista estadístico de las observaciones de la variación de la duración del día. Para ello, se consideran las series de datos observacionales obtenidos a partir de registros astro-geodésicos proporcionados por el *HM Nautical Almanac Office* y el *Servicio Internacional de Rotación de la Tierra y Sistemas de Referencia*, y se lleva a cabo un estudio con el objetivo de obtener el ajuste que modelice de forma más coherente las observaciones de cada serie y explique la variabilidad de la variable. Los resultados evidencian que modelos sencillos, de tipo polinómico cuadrático, permiten describir de manera fiable las observaciones anteriores al s. XIX, necesitando modelos más complejos de tipo splines para registros posteriores.

Palabras y frases clave: Astro-geodésia, Duración del día, Movimiento del polo, Regresión polinómica, Splines.

1. INTRODUCCIÓN

Desde la antigüedad hasta nuestros días, el movimiento de la Tierra ha sido estudiado y analizado en profundidad, especialmente con la llegada de las nuevas tecnologías en el siglo XX. Como se sabe en la actualidad, la Tierra no es una esfera perfecta, sino que es un elipsoide de forma irregular y achatada por los polos, que recibe una serie de acciones gravitatorias tanto del Sol como de la Luna, haciendo que el movimiento de rotación sea relativamente complejo, con variaciones en la velocidad angular de rotación y en la dirección del eje terrestre (Moritz, 1984).

La discusión sobre la evidente variación de la rotación de la Tierra, ha ocupado el interés de astrónomos, geodestas y geofísicos durante más de un siglo y ha incluido el estudio, tanto de una serie de observaciones de telescopio óptico que se remontan a unos 300 años, como de distintos registros históricos de eclipses lunares y solares, y ocultaciones planetarias (Lambeck, 1988). A las técnicas tradicionales de medición hay que añadir los nuevos métodos derivados de los desarrollos tecnológicos orientados al espacio de las últimas décadas.

A partir de las observaciones del tránsito de las estrellas a través del meridiano, han sido descubiertas variaciones periódicas en la duración del día con una precisión que llega a 1 milisegundo en los años cincuenta. Esta tasa de cambio (variación) en la duración del día en relación con un día estándar de 86400 segundos, se denotada de forma genérica como *lod* (de la expresión *length of day*). Además, se han identificado perturbaciones tanto anuales, semi-anuales como diarias, que han ocasionado la introducción del tiempo atómico para la mejora de la determinación del tiempo universal.

El objetivo principal de este trabajo es mostrar y ampliar algunos de los resultados ofrecidos en Junguito et al. (2023), analizando las series de la variable *lod* de interés desde un punto de vista estadístico con el software R (version 4.3.0). Para ello, se considera como punto de referencia el análisis realizado por Sthephenson et al. (2016) and Morrison et al. (2019) y se estudian diferentes tipos de ajustes (lineal, polinómico, splines) para la correcta explicación de la variabilidad de los datos.

2. ANÁLISIS DE LAS SERIES *lod*

Los registros iniciales de eclipses lunares y solares observados desde 700 a.C. hasta 1600 d.C., originarios de civilizaciones antiguas y medievales, han sido considerados por su utilidad para responder a cuestiones acerca de la amplia variabilidad de la rotación de la Tierra. En la Figura 1 se muestran los datos de la variable *lod* de estudio, observada en segundos (s), a lo largo de dicho periodo. Como se podría prever, debido a la precisión de las mediciones, en las civilizaciones más antiguas se tienen registros de una mayor variación de la duración del día en comparación con los datos de las civilizaciones establecidas en la Edad Media.

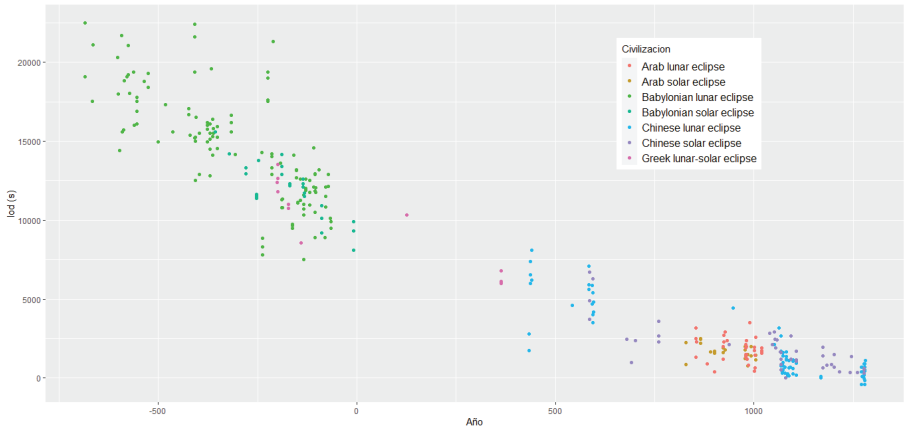


Figura 1: *Variación de la duración del día* registrada por civilizaciones antiguas.

Desde una perspectiva más actual, para el análisis de la variable objetivo de estudio, se utilizan los datos proporcionados por el *HM Nautical Almanac Office* (HMNAO), grupo multidisciplinar de especialistas de la Oficina Hidrográfica del Reino Unido, cuyo propósito es el de proporcionar soluciones a los problemas astronómicos y de navegación celeste. Igualmente se recogen datos del *Servicio Internacional de Rotación de la Tierra y Sistemas de Referencia* (IERS), grupo fundado por la Unión Astronómica Internacional y la Unión de Geodesia y Geofísica Internacional.

En el estudio de las series ofrecidas por HMNAO (Figura 2), se consideran las parejas de valores (X_i, Y_i) con $i = 1, \ldots, n$, donde X_i señala el año de registro e Y_i indica el valor de la variable *lod* de interés, observada en milisegundos (ms). Se pretende encontrar la función que mejor se ajusta al conjunto de datos experimentales con el fin de observar la tendencia, así como plantear predicciones relacionadas con el periodo de estudio. Para ello, se ha utilizado como base la aproximación por mínimos cuadrados, tratando de realizar un ajuste polinómico clásico:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \epsilon_i$$

siendo k el grado de polinomio y ϵ_i el error cometido (residuo). Se comprueba si los coeficientes del polinomio son significativos en el modelo y se evalúa la bondad de ajuste a través del coeficiente de determinación ajustado. Para este análisis, se utiliza la función `lm()` de R que permite incorporar los términos polinómicos con $k \geq 2$ en el modelo (Pardo-Fernández, 2023). Se muestran los resultados relevantes, omitiendo en este trabajo un estudio exhaustivo de residuos.

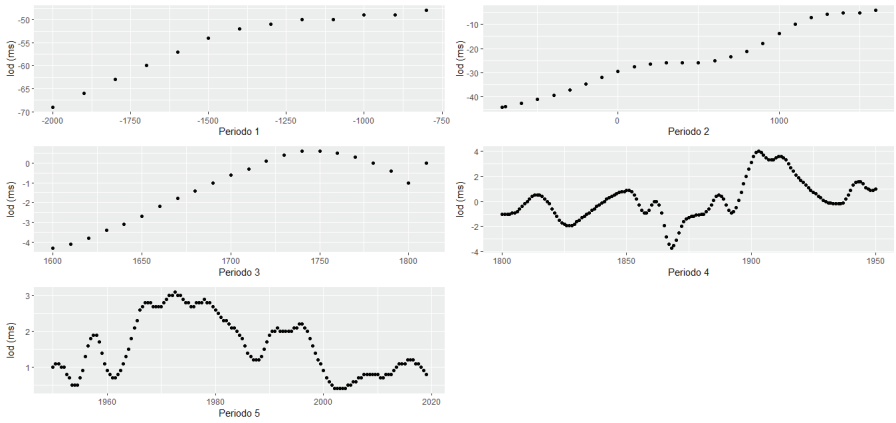


Figura 2: Series de la *variación de la duración del día* desde -2000 hasta 2000.

2.1 Series HMNAO

A continuación se ofrecen las salidas obtenidas para cada uno de los periodos registrados. Se ha planteado en las primera series (Periodo 1, Periodo 2, Periodo 3) un ajuste cuadrático. En el caso de las últimas series (Periodo 4 y Periodo 5) se ha realizado un ajuste cúbico, debido a que el modelo cuadrático no explica de manera adecuada la variabilidad de los datos.

▷ Periodo 1 $(-2000 : -720)$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6.290e+01  2.447e+00 -25.705 1.82e-10 ***
data1$Year      -3.170e-02  3.665e-03  -8.649 5.91e-06 ***
I(data1$Year^2) -1.748e-05  1.300e-06 -13.450 9.93e-08 ***
--
Adjusted R-squared:  0.9933. F-statistic: 891.2 on 2 and 10 DF,  p-value: 5.405e-12
```

▷ Periodo 2 $(-720 : 1600)$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.205e+01  5.963e-01 -53.748  < 2e-16 ***
data2$Year       1.686e-02  9.945e-04  16.951  4.1e-14 ***
I(data2$Year^2)  1.021e-06  9.609e-07   1.062      0.3
---
Adjusted R-squared:  0.9728. F-statistic: 429.4 on 2 and 22 DF,  p-value: < 2.2e-16
```

▷ Periodo 3 $(1600 : 1800)$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6.113e+02  6.924e+01  -8.829 3.76e-08 ***
data3$Year       6.944e-01  8.130e-02   8.541 6.26e-08 ***
I(data3$Year^2) -1.971e-04  2.384e-05  -8.270 1.02e-07 ***
---
Adjusted R-squared:  0.9411. F-statistic: 168.9 on 2 and 19 DF,  p-value: 7.938e-13
```

Los resultados para el Periodo 1, Periodo 2 y Periodo 3, muestran que los términos del ajuste cuadrático son significativos, salvo en el Periodo 2, donde un modelo lineal podría ser suficiente para explicar la tendencia. Así mismo, se aprecia en la Figura 3 que la calidad del ajuste es muy bueno en los 3 casos, obteniéndose que el porcentaje de la variabilidad de la variable *lod* explicado por el modelo es del $99.33\%, 97.3\%$ y 94.11%, respectivamente.

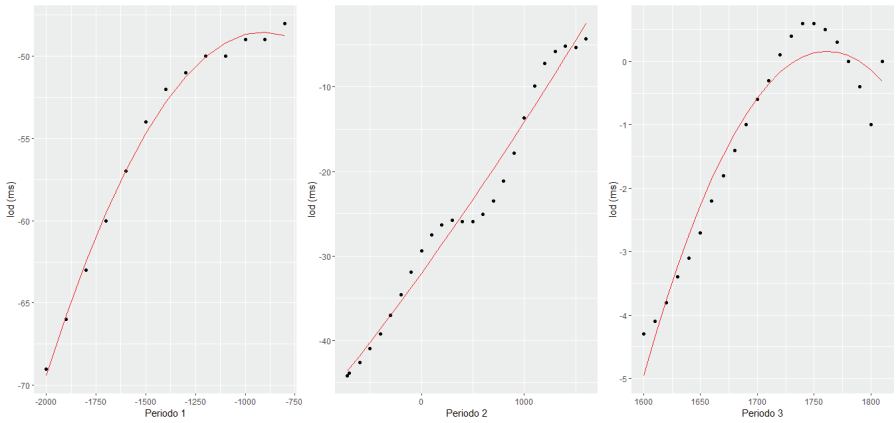


Figura 3: *Variación de la duración del día* registrada hasta 1800.

▷ Periodo 4 (1800 : 1950)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.006e+04  1.098e+04   4.559 1.07e-05 ***
data4$Year      -8.008e+01  1.758e+01  -4.556 1.09e-05 ***
I(data4$Year^2)  4.267e-02  9.377e-03   4.551 1.11e-05 ***
I(data4$Year^3) -7.576e-06  1.667e-06  -4.545 1.14e-05 ***
---
Adjusted R-squared:  0.3223. F-statistic: 24.78 on 3 and 147 DF,  p-value: 4.938e-13
```

▷ Periodo 5 (1950 : 2019)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.492e+05  5.064e+04  -8.871 3.84e-15 ***
data5$Year       6.765e+02  7.656e+01   8.836 4.68e-15 ***
I(data5$Year^2) -3.396e-01  3.858e-02  -8.801 5.70e-15 ***
I(data5$Year^3)  5.681e-05  6.481e-06   8.766 6.96e-15 ***
---
Adjusted R-squared:  0.6491. F-statistic: 86.08 on 3 and 135 DF,  p-value: < 2.2e-16
```

Los resultados para el Periodo 4 y Periodo 5 muestran que en ambas series de datos los términos del modelo cúbico son significativos, pero la calidad del ajuste no es aceptable. En estos casos, el coeficiente de determinación ajustado es del 32.23% y 64.91%, respectivamente, debiendo modelizar estas series de una forma más adecuada.

La estrategia a seguir, será realizar un ajuste por splines de tipo cúbico para dar con un modelo más robusto (De Boor, 1978). Se considera:

$$Y_i = \alpha_1 B_1(X_i) + \cdots + \alpha_j B_j(X_i)$$

donde j indica el número de bases, $\{a_1, \ldots, a_j\}$ los parámetros desconocidos y $\{B_1, \ldots, B_j\}$ base de funciones que depende de la posición de los nodos establecida. Consideramos en R, la función `lm()` junto con el función `bs()` que genera la base mencionada de tipo cúbico (Wood, 2017), incorporando los nodos `knots` según lo indicado por Stephenson et al. (2016).

▷ Periodo 4 (1800 : 1950)

```
> lm(formula = data4$lod ~ bs(data4$Year, knots = seq(1800, 1900, 5)))

Adjusted R-squared:  0.9187. F-statistic: 74.69 on 23 and 127 DF,  p-value: < 2.2e-16
```

▷ Periodo 5 (1950 : 2019)

```
> lm(formula = data5$lod ~ bs(data5$Year, knots = seq(1900, 2016, 3)))

Adjusted R-squared:  0.9949. F-statistic:  1072 on 25 and 113 DF,  p-value: < 2.2e-16
```

Las salidas demuestran que, en esta ocasión, el ajuste no solo es adecuado sino que la proporción explicada de la variabilidad de la variable *lod* es muy alta (Figura 4). En estos casos, el coeficiente de determinación ajustado llega a ser del 91.87% y 99.49%, respectivamente.

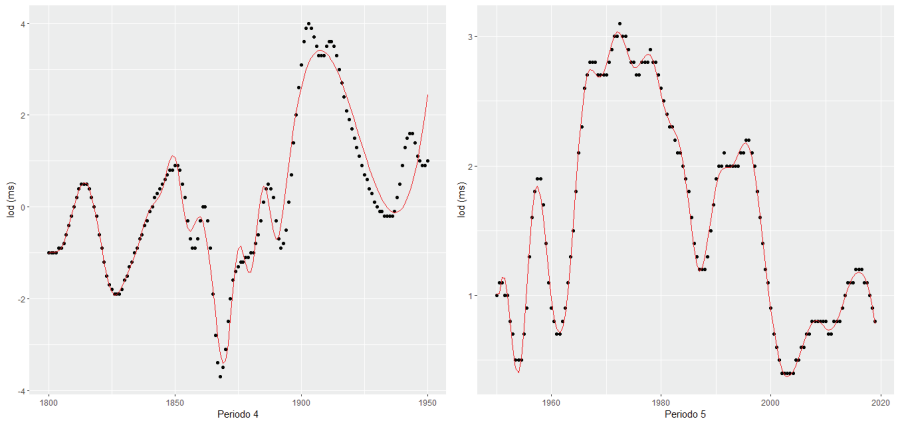


Figura 4: *Variación de duración del día* modelizada por splines.

2.2 Series IERS

Adicionalmente, se puede realizar una comparativa entre las series registradas y estudiadas en el punto anterior y los datos proporcionados por el IERS. Inicialmente, se anotan los registros establecidos desde el año 1950 al año 2010 con un intervalo de mediciones similar (cada año o año y medio). Se puede observar en la Figura 5 izda) que en ambos casos los datos obtenidos son muy similares, pudiendo ajustar un modelo cúbico por splines, aunque no con toda la precisión requerida. Suponiendo las series de forma independiente, se obtienen las medidas resumen descriptivas que indican una dispersión similar, con coeficientes de variación de 0.4917 para los datos del HMNAO, y 0.4644 para los datos del IERS.

Centrándose en los datos propios del s. XXI que contiene el IERS (desde el año 2000 al año 2023) se puede apreciar en la Figura 5 dcha) que las observaciones se toman con una mayor frecuencia temporal (de forma diaria, con un total de 365 registros por año), existiendo una mayor variabilidad de la variable *lod* de interés (coeficiente de variación de 1.0608).

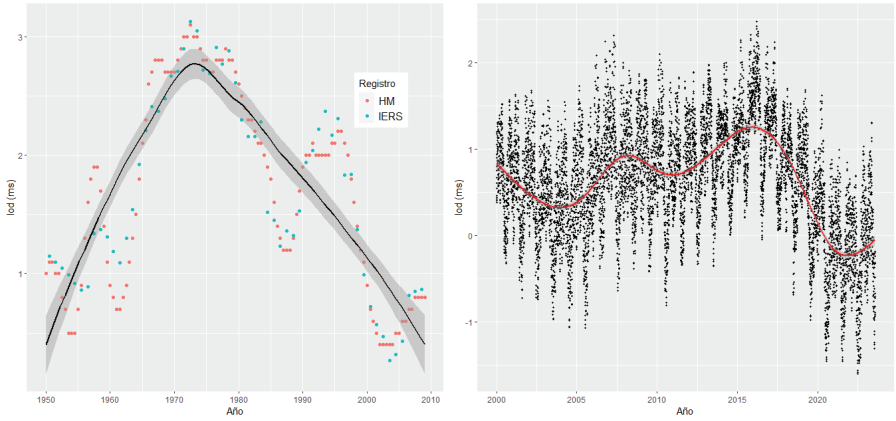


Figura 5: izda) Comparativa entre series IERS y HMNAO; dcha) Serie del s. XXI del IERS.

3. CONCLUSIONES

Este trabajo presenta un estudio de la variación de la duración del día (*lod*) desde la perspectiva observacional. El análisis estadístico que se ha realizado permite modelizar las series de datos registradas desde la antigüedad hasta la actualidad. Los resultados obtenidos indican que las observaciones contempladas hasta el s. XIX, se pueden ajustar a un modelo polinómico cuadrático, de manera adecuada y fiable. Sin embargo, los datos observados a partir del s. XIX, no permiten un ajuste polinómico clásico, sino que se ajustan mejor a un modelo tipo cúbico spline. Si se realiza la comparación de los registros entre las dos grandes agencias de investigación en geodesia IERS y HMNAO se observa que los registros del periodo 1950 - 2010 son similares, pudiendo ajustar un modelo cúbico spline. Además, los registros completos del s. XXI del IERS indican una mayor dipersión de la variable de interés en comparación con periodos previos.

REFERENCIAS

De Boor, C. (1978) A practical guide to splines (Vol. 27). Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

HM Nautical Almanac Office.
`https://www.gov.uk/government/organisations/hm-nautical-almanac-office`.

International Earth Rotation and Reference Systems Service. `https://www.iers.org/IERS/EN/Home/home_node.html`.

Junguito, I, Folgueira, M. and Álvarez M. (2023) TFG: Análisis de las series de la variación de la duración del día. Repositorio CUD-ENM `http://calderon.cud.uvigo.es/handle/123456789/690`.

Lambeck, K. (1988) The Earths variable rotation: some geophysical causes. Astrophysics and Space Science Library, vol 154. Springer, Dordrecht.

Moritz, H. (1984) Rotación de la Tierra. Instituto de Astronomía y Geodesia. N 136. CSIC.

Morrison, L. V., Stephenson, F. R., Hohenkerk, C. Y. and Zawilski, M. (2021) Addendum 2020 to "Measurement of the Earths rotation: 720 BC to AD 2015". Proceedings of the Royal Society A, 477: 20200776.

Pardo Fernández, J. C. (2023) Bioestatística para a Enxeñaría Biomédica. Edición: Universidade de Vigo.

R Core Team (2023) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/.`

Stephenson, F. R., Morrison, L. V. and Hohenkerk, C. Y. (2016) Measurement of the Earth's rotation: 720 BC to AD 2015. Proceedings of the Royal Society A, 472: 2016040.

Wood, S. N. (2017) Generalized additive models: an introduction with R, Second Edition. Chapman & Hall/CRC Texts in Statistical Science.

25º ANIVERSARIO DO GRUPO SiDOR (1998-2023)

Jacobo de Uña Álvarez[1], Gloria Fiestras Janeiro[1] e Javier Roca Pardiñas[1]

[1] Grupo SiDOR, Universidade de Vigo

RESUMO

Durante o presente ano 2023 o grupo de investigación Inferencia Estatística, Decisión e Investigación Operativa (SiDOR) da Universidade de Vigo (http://sidor.uvigo.es) celebra o seu 25º aniversario. Fundado en 1998, o grupo SiDOR asentouse ao longo destes anos como un grupo de referencia na área de Estatística e Investigación Operativa, tanto a nivel autonómico como nacional e internacional. Non é casualidade que a Xunta de Galicia inclúa a SiDOR dende o ano 2008 no seu catálogo de unidades de investigación excelentes, o cal dota ao grupo dunha importante financiación competitiva estrutural. Os catro piares da actividade do SiDOR son a formación, a investigación, a transferencia e a divulgación. Nesta contribución os coordinadores do SiDOR fan memoria dos 25 anos de actividade do grupo, enumerando os fitos máis salientables na súa traxectoria, e reflexionan sobre os principais retos que deberá afrontar o SiDOR no futuro próximo.

Palabras e frases chave: Divulgación, Estatística e Investigación Operativa, Formación, Investigación, Transferencia

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

ANÁLISE MULTIVARIANTE APLICADA Á LIBERACIÓN DE FIBRAS TÉXTILES

Mercedes Pereira[1], Laura S. Vázquez[1], Ana-María Díaz-Díaz[1], Salvador Naya[2], Jorge López-Beceiro[1]

[1] Centro de Investigacións en Tecnoloxías Navais e Industriais, Universidade da Coruña, Campus Industrial de Ferrol, Ferrol

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

RESUMO

A contaminación producida polos materiais téxtiles leva sendo motivo de preocupación nos últimos tempos debido ás microfibras que desprenden durante o seu uso diario e mantemento, resultando nunha acumulación de microplásticos no medio ambiente, especialmente nas masas de auga. Unha adecuada comprensión do mecanismo polo cal as fibras son liberadas dos téxtiles é un reto no que o papel da matemática pode aportar unha contribución substancial. Con esta perspectiva, neste traballo estúdanse diferentes téxtiles a través da medición de diferentes variables dimensionais, mecánicas e de composición, ademais daquelas relacionadas coa liberación de fibras. Mediante técnicas estatísticas de análise multivariante, como é o caso da Análise de Compoñentes Principais, búscase identificar a relación de esta liberación de fibras coas características dimensionais, mecánicas e composicionais dos tecidos.

Palabras e frases chave: Análise exploratoria, Análise de Compoñentes Principais, Téxtil, Análise Termomecánica, Liberación de fibras

1. INTRODUCCIÓN

A complexidade dos materiais téxtiles en canto a estrutura e composición engloba moitas variables, o cal dificulta a visualización de datos e, en consecuencia, o seu estudo. Variacións na natureza das fibras utilizadas nun téxtil, así como a súa disposición en fíos, que á súa vez serán tecidos de diversas maneiras desemboca en diferenzas de diversa natureza entre téxtiles.

Esta complexidade de variables a ter en conta no estudo dos materiais téxtiles dificulta enormemente a resolución de problemas como a liberación de fibras nas prendas no seu uso e mantemento. A imposibilidade de visualización de datos require necesariamente de técnicas de redución de dimensións como é a Análise de Compoñentes Principais (PCA).

Neste estudo analízanse mediante PCA algunhas propiedades de diferentes tecidos de calada. Inclúense propiedades dimensionais tanto a nivel de fibra como de fío e tea, xunto coa súa composición en fibras, ademais de certos comportamentos dinámicos e termomecánicos. Por último, a análise inclúe a capacidade de liberación de fibras tanto en ambiente seco coma húmido.

2. MATERIAIS E PROCEDEMENTO

Neste estudo utilizáronse 15 tecidos comerciais de calada compostos por mesturas das seguintes fibras: poliéster, acrílico, elastano, viscosa, algodón e lá. O ligamento de cinco deles é tafetá, mentres que outras sete son sargas e tres son de satén. As súas propiedades dimensionais e de composición foron en parte medidas con métodos ópticos e gravimétricos, e completados con información proporcionada polo provedor.

Todas as teas foron sometidas a ensaios termomecánicos no instrumento Rheometrics DMTA-IV. O primeiro consiste nunha rampla de temperatura dende a temperatura ambiente a 60 °C, mantendo unha tensión constante sobre os tecidos e medindo a súa deformación lineal a 50 °C. O segundo tipo de ensaio sobre as mostras é un barrido de frecuencias comprendido entre 100 e 0,1 Hz, que permite medir o módulo elástico dos tecidos.

Por último, leváronse a cabo experimentos de liberación de fibras en seco utilizando unha máquina Random Tumble Pilling Tester de SD Atlas, segundo unha adaptación da norma ISO 1295-3:2020. O provedor compartiu a información de liberación en ambiente húmido para experimentos nunha lavadora de laboratorio a 30 °C durante 40 min.

3. ANÁLISE DE DATOS

A análise de datos mediante PCA realizouse utilizando o software `R`, en concreto `RStudio`, a través da función `prcomp()` do paquete `Stats`. As 13 variables estudadas móstranse na Táboa 1, onde tamén se indica, por motivos de contexto, o seu tipo: dimensional, sexa a nivel de fibra, fío ou tea; relacionadas coa composición, sumando a porcentaxe total de fibras artificiais dentro das mesturas con fibras naturais; tamén as variables indicativas do comportamento mecánico; e aquelas relacionadas coa degradación, en concreto coa liberación de fibras, tanto en ambiente húmido como en seco.

Nos experimentos en ambiente seco se mediu a perda de fibras ao longo do tempo, polo que foi posible observar unha tendencia potencial, na que unha primeira gran liberación de fibras é seguida por unha estabilización no tempo. Desta maneira, se distinguen dúas propiedades en ambiente seco: a perda inicial e a capacidade de continuar liberando fibras a tempos maiores.

Táboa 1: Variables estudadas mediante PCA

Tipo de propiedade	**Propiedade**	**Nome curto**
Composición	Porcentaxe de fibra artificial	Fibras_artificiais
Dimensional – fibra	Diámetro da fibra	Diametro_fibra
Dimensional – fibra	Lonxitude da fibra	Lonxitude_fibra
Dimensional – fío	Densidade lineal do fío	Titulo_fio
Dimensional – tea	Coeficiente de ligadura	Coef_ligadura
Dimensional – tea	Factor de cobertura	Factor_cobertura
Dimensional – tea	Espesor da tea	Espesor_tea
Dimensional – tea	Densidade de área da tea	Gramaxe
Mecánica	Rixidez da tea	Modulo
Mecánica	Deformación a 50 °C	Deformación_50C
Liberación – Húmido	Perda de fibras en lavadora	Perda_mollado
Liberación – Seco	Perda inicial de fibras en seco	Perda_inicial
Liberación – Seco	Perda de fibras a longo prazo en seco	Perda_no_tempo

A redución de dimensións mediante PCA e a súa posterior análise revelou a necesidade de ter en conta as catro primeiras compoñentes principais (PCs), que representan o 77,1 % da variabilidade total dos datos. Estúdanse, por tanto PC1, PC2, PC3 e PC4, que explican un %, %, % e % da varianza, respectivamente. As variables con máis peso nas tres primeiras compoñentes principais son aquelas relacionadas cas dimensións das estruturas téxtiles xunto ca perda de fibras, mentres que PC4 está máis ligada ás propiedades mecánicas.

Como resultado da análise das dúas primeiras compoñentes principais, utilizando ferramentas gráficas como o biplot, atopouse que a perda inicial de fibras nos experimentos en seco está estreitamente relacionada cunha baixa proporción de fibras artificiais, e cunha lonxitude de fibra curta. En contraposición, as fibras artificiais de gran lonxitude tenden a continuar liberando fibras a longo prazo, despois desa perda inicial. Variables relacionadas cas dimensións da tea e os fíos resultan en pouca ou nula relación coa liberación de fibras en seco, mentres que a perda en mollado se observa na mesma dirección que a deformación mecánica medida a 50 °C.

Por outra banda, ao estudar as compoñentes principais 3 e 4, atópase que en experimentos de lavado cobran importancia as propiedades termomecánicas, xa que unha deformación a 50 °C máis acusada se relaciona coa perda de fibras nestas condicións. Un maior módulo elástico tamén se relaciona con esta perda en mollado, mais tendo en conta PC1 e PC2, onde non se observa en absoluto esta relación, non se pode extraer a conclusión de que estas dúas propiedades estean fortemente relacionadas.

Por último, a gráfica que representa PC1 fronte a PC3 reforza esta relación entre a deformación a 50 °C e a perda de fibras en mollado, sendo asociada tamén ca perda inicial en seco. Esta última variable aparece fortemente relacionada ca densidade lineal dos fíos, mentres que a perda a longo prazo vese favorecida por un menor factor de cobertura.

4. CONCLUSIÓNS

Deste estudo se extrae que a liberación de fibras en distintos ambientes de humidade dependen de propiedades con orixes independentes. Con respecto á perda durante o uso e secado mecánico dos téxtiles, simulados nos experimentos en seco, atópase que unha perda inicial está promovida por fibras naturais curtas, mentres que as fibras artificiais longas tenderán a continuar liberando microplásticos despois desa primeira perda. Durante o lavado mecánico dos téxtiles serán de importancia as propiedades mecánicas dos mesmos, xunto coa estrutura de teas e fíos.

REFERENCIAS

European Comitee for Standarization (2021) Textiles. Determination of fabric propensity to surface pilling, fuzzing or matting. Part 3: Random tumble pilling method (ISO 12945-3:2020)

R Core Team (2022) R: A language and environment for statistical analysis. Version 2022.07.2+576. URL: https://www.R-project.org/

Wehrens, R. (2011) Chemometrics with R (Vol. 3). New York: Springer.

A generalized additive model (GAM) approach to regression and variable selection of geographic data.

Francisco de Asis Lopez[1],*Javier Roca-Pardinas*[2], *Celestino Ordonez.*[3].

[1]franciscoasis.lopez@uvigo.es,Department of Statistics and Operational Research. University of Vigo
[2]roca@uvigo.es, Department of Statistics and Operational Research. University of Vigo
[3]cgalan.uniovi@gmail.com, Department of Mining Exploitation and Prospecting. University of Oviedo

Abstract

Regression models for spatial data have attracted the attention of researchers from different fields given their widespread application. In this work we analyze the utility of generalized additive models (GAMs) as a regression method with spatially-dependent coefficients and compared them with the geographically weighted regression (GWR), a popular method that performs a local linear regression in the vicinity of each point. The comparison was carried out using both simulated and real data. The results of the comparison in the simulated data shows some advantages of GAM for spatial data over GWR. The comparison using real data shows quite similar results for both methods, although the root mean square in GAM was slightly lower for the GAM approach.

We also propose a bootstrapping-based method to test for spatial heterogeneity of the models coefficients and a stepwise procedure to select the significant covariates using BIC (Bayesian information criterion). The application of these two procedures to simulated data was successful.

Xornadas de innovación docente na estatística e investigación de operacións

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26-28 de outubro de 2023

TÉCNICAS DE INNOVACIÓN DOCENTE MEDIANTE HERRAMIENTAS TIC EN ASIGNATURAS DE ESTADÍSTICA Y ECONOMETRÍA

María-Carmen Sánchez-Sellero[1], Beatriz García-Carro[2]

[1] Profesor Titular de Universidad. Universidade da Coruña, Facultad de Economía y Empresa, Departamento de Economía, Campus de Elviña, A Coruña, España.
[2] Profesor Titular de Universidad. Universidade da Coruña, Facultad de Economía y Empresa, Departamento de Economía, Campus de Elviña, A Coruña, España.

RESUMEN

La innovación en general y la innovación educativa en particular es uno de los grandes retos de nuestra sociedad. La Educación Superior en fechas recientes se mueve en la búsqueda de nuevas metodologías y contenidos en aras a que los indicadores de calidad mejoren. En este proceso, el aprendizaje de los alumnos es uno de los objetivos principales de las instituciones educativas. Las tecnologías de la información y las comunicaciones (TIC) son herramientas útiles en este proceso de innovación educativa. Actitudes proactivas, implicadas y positivas de la comunidad universitaria son también factores necesarios que conducen al progreso y la mejora. En este trabajo se presenta un escueto análisis de estos aspectos, que se completa con un estudio específico de las experiencias docentes en el desarrollo de técnicas de innovación aplicadas en las asignaturas de Estadística y Econometría de la facultad de Economía y Empresa de la Universidade da Coruña (UDC).

Palabras y frases clave: Innovación docente, hoja de cálculo, campus virtual, Estadística.

1. INTRODUCCIÓN

En el entorno actual de cambios continuos y de transformación digital en que estamos inmersos, la innovación es un término que resulta imprescindible. Innovar implica tomar decisiones que permiten introducir cambios, modificar ideas, actitudes, y cultura para el desarrollo de nuevas prácticas, que conlleven o aporten mejoras. El concepto de "innovación" está asociado a la creación, percepción y asimilación de algo novedoso (Margalef y Arenas (2006); Macanchí et al. (2020)). En palabras de Macanchí et al. (2020), "Desarrollar una cultura de la innovación se ha convertido en uno de los retos más importantes en la Educación Superior".

Dada la importancia en el ámbito académico y su mayor repercusión en momentos como la situación sanitaria vivida en la COVID-19, se imponen nuevos desafíos para la organización de las universidades y el desarrollo de las actividades docentes.

No solamente se trata de generar cambios en la práctica docente, sino de hacer todo lo posible por incentivar ideas que modifiquen la actuación del profesorado y del alumnado, lo cual supone transformaciones en los procesos didácticos. Ello implica una renovación sustancial en el proceso de aprendizaje, en cuanto a métodos, contenidos y materiales que conduzcan a una mejora de los resultados académicos de los alumnos.

Por otro lado, las tecnologías de la información y las comunicaciones (TIC) aportan diferentes modelos de actuación, nuevas formas y escenarios para el desarrollo de esos procesos docentes; también conducen a cambios significativos en los roles del profesor y del alumno. Todo ello implica, a su vez, cambios en los cánones de enseñanza-aprendizaje hacia un modelo más flexible (Salinas, 2004). La política institucional debe involucrar a todos y sensibilizar a los miembros de la comunidad universitaria. La implementación de sistemas de apoyo a profesores y alumnos, así como la

infraestructura tecnológica basada en líneas estratégicas y un plan fundamentado, garantizan el éxito y la sostenibilidad de los procesos educativos (Vidal et al., 2022).

El objetivo último de la Educación Superior es dotar a los estudiantes de una formación académica de calidad que esté al mismo tiempo adaptada a las necesidades del mercado laboral. Es incuestionable que en los últimos años las TIC y su vertiginoso desarrollo han generado nuevas necesidades en la formación universitaria, y al mismo tiempo estas nuevas tecnologías se han convertido en un motor de cambio e innovación dentro de la Universidad. Por este motivo, creemos que en la docencia universitaria es necesaria la innovación, y dicha innovación debe tener en las TIC su principal sustento (Quiroga et al., 2019).

Todos los docentes debemos dedicar esfuerzos y tiempo a mejorar la calidad de nuestra docencia, y es nuestra responsabilidad utilizar metodologías y herramientas de aprendizaje que doten a los estudiantes de destrezas, habilidades y conocimientos adecuados. La mejora va de la mano de la innovación, algunos autores consideran que la clave del éxito de la innovación docente es que el profesorado perciba la innovación como un elemento necesario, fácil, útil y eficiente (Marqués, 2015). Asimismo, se argumenta que la innovación que favorece a la enseñanza, requiere de una formación tecnológica permanente (Flores y Meléndez, 2021).

En el ámbito concreto de las asignaturas de Estadística, que es lo que desarrollaremos en el epígrafe siguiente, la implementación de herramientas educativas de TIC permite conocer y explorar las aplicaciones, y de este modo profundizar en el aprendizaje de estos temas, generando una cultura estadística beneficiosa para situaciones cotidianas (Ramírez y Rodríguez, 2023).

2. EXPERIENCIAS DOCENTES INNOVADORAS EN ASIGNATURAS DE ESTADÍSTICA Y ECONOMETRÍA

En este trabajo queremos mostrar las herramientas docentes que hemos utilizado en las clases de Estadística y Econometría. Estas asignaturas las impartimos en los grados universitarios de la Facultad de Economía y Empresa de la UDC. Nuestra forma de trabajar es testar las herramientas en una asignatura concreta. Si los docentes vemos sus beneficios en el aprendizaje del alumnado, posteriormente las implementamos en el resto de las asignaturas de las que somos responsables.

En el curso 2020/2021 la asignatura “Estadística I” tenía estipulada en su guía docente un porcentaje de la nota correspondiente a las prácticas a través de TIC. En el transcurso de la crisis sanitaria del COVID-19, estas prácticas no se pudieron realizar de modo presencial, dado que en el aula de informática no se podían mantener las distancias de seguridad. De este modo, las prácticas se desarrollaron de la siguiente manera:

- Primero, se prepararon varios seminarios en formato virtual a través de la aplicación MS Teams, con el fin de explicar en una hoja de cálculo (Excel) la aplicación práctica de los conceptos estudiados en la primera parte de la asignatura, que se engloban en la denominada Estadística Descriptiva. Para esto, se seleccionaron como base de datos varias series económicas reales obtenidas de la web del INE (Instituto Nacional de Estadística).
- Segundo, dado que estas prácticas son evaluables, las profesoras creamos una tarea en el Campus Virtual para cada grupo de clase, planificada para que los alumnos la realizasen el día y hora estipulados. Con ello, se pretendía que los alumnos en grupos de 2-3 personas reprodujesen con los datos que le tocasen, lo aprendido en los seminarios.
- Tercero, el día de la prueba los estudiantes se conectaron a MS Teams para recibir las instrucciones de la profesora. Se añadía otra dificultad, consistente en que los estudiantes debían estar conectados entre ellos, dado el carácter grupal de la tarea. La tarea se explicaba de manera oral y escrita en el Campus Virtual (Moodle). En el Campus Virtual aparecían previamente las indicaciones para realizar la prueba/trabajo de Excel.

En los cursos siguientes, también se realizaron estas prácticas, aunque con pequeñas variaciones, ya que los seminarios y las pruebas volvieron a la presencialidad. Actualmente, los seminarios los desarrollamos indistintamente en el aula de informática o en el aula de clase ya que los alumnos traen sus ordenadores portátiles.

Con estas técnicas, las profesoras combinamos la enseñanza de las TIC, en este caso el uso de una hoja de cálculo (Excel) con el uso de la plataforma del Campus Virtual y la aplicación MS Teams. A su vez, y fruto de esta innovación docente realizada por las profesoras de la materia, se trabajaron varias de las competencias propias de la titulación como son: Utilizar las herramientas básicas de las TIC necesarias para el ejercicio de su profesión, así como saber trabajar en equipo.

Una combinación de herramientas parecida también se ha utilizado en otras asignaturas, como "Estadística II".

La situación vivida tras el COVID-19 ha supuesto una informatización de la actividad docente, y una mejora sustancial en nuestras competencias en tecnología. En concreto, nos vimos en la necesidad de utilizar bancos de preguntas para la evaluación virtual de las asignaturas. Actualmente, estos cuestionarios de preguntas tipo test es una de las actividades más valorada por los alumnos, aunque su uso no necesariamente tiene una finalidad evaluadora. La plataforma Moodle nos permite crear, organizar y editar un banco de preguntas de manera muy fácil. Nosotras dividimos el banco de preguntas en carpetas, una para cada tema de la asignatura y dentro de cada carpeta creamos subcarpetas en función de los epígrafes. El trabajo más pesado y minucioso es elaborar e introducir en el banco muchas y variadas preguntas tipo test, todas ellas con cuatro posibles respuestas, siendo solo una la respuesta correcta. Lo más importante es que el banco esté bien organizado, para que con el paso del tiempo nos resulte fácil encontrar preguntas que se adapten a los contenidos de los epígrafes de nuestro temario.

Posteriormente, con el banco de preguntas generamos cuestionarios, uno para cada tema. Cada cuestionario consta de un amplio abanico de preguntas aleatorias que pueden ser tanto teóricas como prácticas, pero que engloban todos los contenidos del tema. Estos cuestionarios quedan a disposición de los alumnos en Moodle. De manera on-line los alumnos pueden contestar los cuestionarios todas las veces que quieran y se les permite el acceso hasta la realización del examen final. Además, los alumnos pueden conocer de manera inmediata el resultado obtenido, lo que les permite comprobar su nivel de conocimientos. La técnica de los cuestionarios se ha utilizado en las asignaturas de Estadística y Econometría en las que impartimos docencia.

Nuestra experiencia confirma la gran aceptación por parte de los alumnos de estos cuestionarios. En general, los alumnos responden a todos los cuestionarios a lo largo del curso y, además, vuelven a hacer los cuestionarios cuando la fecha del examen final está cerca.

Para rentabilizar el banco de preguntas e incentivar la atención de los alumnos en clase, hemos desarrollado la técnica del cuestionario sorpresa. Una vez a la semana, sin previo aviso, la profesora pide a los alumnos que realicen de manera individual un cuestionario sorpresa de cuatro o cinco preguntas tipo test. El cuestionario se elabora in situ, lo que requiere dominar el contenido del banco de preguntas, y las preguntas se centrarán en los conceptos explicados en las últimas clases. La ventaja de estos cuestionarios generados al momento es su flexibilidad, pues podemos ajustar su contenido y su duración al tiempo disponible de cada clase. Después, estos cuestionarios quedan a disposición del alumnado a lo largo de todo el curso.

Nuestra experiencia sobre esta técnica es muy satisfactoria. Creemos que los cuestionarios sorpresa aportan mucho dinamismo a las clases y aumenta la motivación del alumnado, pues sabe que semanalmente puede demostrar sus conocimientos. Además, nos permite controlar la asistencia a clase y conocer, en función de las respuestas, si los alumnos han entendido lo explicado. Los alumnos valoran muy positivamente estos cuestionarios, dicen que las clases son más divertidas y agradecen poder hacer en casa con más calma estos cuestionarios y así repasar lo que han aprendido en clase.

Con todas estas actividades se contribuye a las siguientes líneas de actuación del GID (Grupo de Innovación Docente) al que pertenecemos las profesoras:

- Línea 1. Metodologías activas en docencia presencial y semipresencial.
- Línea 7. Instrumentos y recursos para la mejora de la organización y el desempeño docente.

3. CONCLUSIONES

En este estudio hemos expuesto nuestras experiencias docentes en el ámbito de las asignaturas de Estadística y Econometría en aquellas actividades en las que las TIC tenían un papel determinante. El especial momento vivido durante la pandemia de COVID-19 de 2020, hizo que por mera necesidad, todos los docentes tuviésemos que mejorar nuestras habilidades tecnológicas. Lo aprendido en ese contexto, nos proporcionó una mejora en la calidad de la docencia y una adaptación a los entornos virtuales y tecnológicos de hoy en día.

En esta exposición nos hemos centrado en explicar lo que hicimos las autoras para reorganizar la docencia de la Estadística durante la pandemia, siendo conscientes de que estas iniciativas pueden coincidir perfectamente con lo realizado por otros docentes universitarios de Estadística e Investigación de Operaciones. En nuestras clases utilizamos dos técnicas: 1) El uso combinado del Campo Virtual, una hoja de cálculo y la aplicación Teams, y 2) El uso de los Cuestionarios confeccionados en el Campus Virtual.

Fruto de esta experiencia podemos concluir que ambas formas de proceder las consideramos ampliamente satisfactorias por varios motivos: 1) Han contado con valoraciones positivas por parte del alumnado; 2) Incrementan la motivación de los alumnos, al visualizar la materia de un modo aplicado mediante la utilización de un programa informático; 3) Facilitan el aprendizaje de los contenidos de la materia; 4) En último término, añaden valor a la asignatura en cuestión.

REFERENCIAS

Flores, L.., & Meléndez, C. (2021). Análisis comparativo del b-learning y e-learning en competencias TIC para la docencia en educación superior. Revista Innova Educación, 3(4), 173-190.

Macanchí, M. L., Bélgica O. C., & Campoverde, M. A. (2020). Innovación educativa, pedagógica y didáctica. Concepciones para la práctica en la Educación Superior. Universidad y Sociedad, 12(1), 396-403.

Margalef, L., & Arenas, A. (2006). ¿Qué entendemos por innovación educativa? A propósito del desarrollo curricular. Perspectiva Educacional, Formación de Profesores.47, 13-31.

Marqués, P. (2015). Cómo innovar en los centros docentes. DIM: Didáctica, Innovación y Multimedia.

Quiroga, L. P., Jaramillo, S., & Vanegas, O. L. (2019). Ventajas y desventajas de las TIC en la Educación "Desde la primera infancia hasta la Educación Superior". Revista Educación y Pensamiento, 26(26), 77-85.

Ramírez,, L., & Rodríguez, J. A. (2023). Implementation of Technological Tools Teaching Probability and Statistics: Systematic Review. EDU REVIEW. International Education and Learning Review / Revista Internacional de Educación y Aprendizaje, 11(2), 155–171.

Salinas, J. (2004). Innovación docente y uso de las TIC en la enseñanza universitaria. Revista Universidad y Sociedad del Conocimiento, 1(1).

Vidal, M. J., Miralles, E. D. L. Á., Morales, I. D. R., & Gari, M. (2022). Innovación educativa. Educación Médica Superior, 36(3).

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

UNA BUENA FORMACIÓN ESTADÍSTICA EN SECUNDARIA Y BACHILLERATO MEDIANTE PROYECTOS

Jesús Manuel Díaz López[1]

[1]IES Plurilingüe Antón Losada Diéguez. Consellería de Cultura, Educación e Ordenación Universitaria. Xunta de Galicia

RESUMEN

Mi intervención se centrará en la exposición de un par de ejemplos de proyectos de investigación estadística realizados en el aula. Expondré como realizo estos proyectos con el alumnado destacando su relación con nuestro día a día:

- Proyecto 1: **¿El Pan que comemos es saludable?**
  El objetivo de este proyecto fue estudiar la calidad del pan que comemos.
  Basándonos en la "Ley del Pan" del 2019, la cual recoge ciertos parámetros de calidad del pan, analizamos distintas variables para determinar si el pan que comemos es saludable. Además estudiamos relaciones entre el pan saludable y otras variables como por ejemplo que esté hecho con masa madre, ...).
  Este proyecto ha servido para conectar la realidad y la actualidad con el aula.
  Proyecto realizado con alumnos de 4º de ESO.
- Proyecto 2: **Dieta y rendimiento académico**
  El objetivo de este proyecto fue estudiar el tipo de dieta que sigue el alumnado y su posible relación con su rendimiento académico.
  Tomando la nota media del último curso académico o etapa educativa se ha analizado si existe alguna relación entre el rendimiento académico con el tipo de dieta que sigue el alumnado (atlántica, mediterránea, mixta o basada en alimentos procesados). Además se ha analizado la percepción que el alumnado, de la muestra, tiene referente a su alimentación (saludable, poco saludable o no saludable) y se ha comparado con sus respuestas sobre los alimentos que consume habitualmente.
  Proyecto realizado con alumnos de 1º de BAC.

Palabras y frases clave: proyecto, estadística, secundaria, bachillerato, pan, dieta.

XVI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 26 - 28 de outubro de 2023

INTRODUCIÓN Á INVESTIGACIÓN EN ESTATÍSTICA NO STEMBACH

Joaquín García Lamela[1]

[1]Profesor de Matemáticas e coordinador do STEMbach do IES María Casares (Oleiros)

RESUMO

A participación do alumnado no programa STEMBach supón unha oportunidade de traballar e investigar dedicándolle máis tempo do que habitualmente é posible. No IES María Casares decidimos centrarnos na estatística e na análise de datos, buscando diferentes enfoques para conseguir a súa interese nestes temas e rematando o proceso coa redacción e exposición do traballo estatístico realizado sobre un tema escollido por eles mesmos.

Palabras e frases chave: STEM, bacharelato, análise de datos, estatística, folla de cálculo.

1. INTRODUCIÓN

O STEMbach é un bacharelato de excelencia en Ciencias e Tecnoloxía que se estableceu por primeira vez no curso 2018/19. A finalidade do programa é que o alumnado poida afondar na competencia STEM mediante o coñecemento e a realización de traballos de investigación e do método científico.

Tendo en conta esta finalidade, o programa permite que cada centro organice a súa propia materia extracurricular. No caso do IES María Casares centrámonos na análise de datos mediante estatística, co que relacionamos os traballos que realizará o alumnado coa expresión *Big Data* que todos eles escoitaron algunha vez aínda que posiblemente sen ter moi claro o seu significado.

Estes traballos van ser realizados por alumnado de 1º de Bacharelato cuxos coñecementos estatísticos son bastante limitados, polo que hai que introducir conceptos previos e darlles ferramentas para poder facer a análise, mais a vantaxe da materia STEMbach é que é moi flexible e permite tratar a teoría con tempo, adecuándoa ás características de cada alumno. Estas ferramentas e conceptos servirán ao alumnado para realizar o seu estudo.

2. CONCEPTOS TEÓRICOS

Antes de explicar ao alumnado como facer a investigación dedicamos varias clases a que aprendan por unha banda o manexo dunha folla de cálculo e por outra as principais ferramentas estatísticas que poderán precisar.

O uso da folla de cálculo non só é básico en moitos traballos na actualidade, senón que é un programa que todo o alumnado terá nos seus computadores pero que pode ser empregada coma unha potente ferramenta estatística. Temos que dedicar varias clases para que coñezan o seu uso porque, a pesar do que eles adoitan pensar, apenas saben que facer con ela, máis aló de cambiar o aspecto da folla. É unha oportunidade para introducir utilidades do programa como ordenar datos, o uso de funcións (non necesariamente matemáticas) ou resaltar celas con datos salientables.

Unha vez coñecido o uso básico dunha folla de cálculo pasamos a explicar algúns conceptos básicos de estatística, sempre apoiándonos nas funcións do programa. A principal dificultade que se atopa neste momento é introducir estes conceptos de forma que o alumnado se decate da importancia que na actualidade ten a estatística en case calquera disciplina e se sinta atraído por ela. É moi habitual que, a pesar de traballar exclusivamente cos computadores e coa folla de cálculo, o alumnado os siga a ver como un árido conxunto de números e operacións.

Unha solución que pensamos é desenvolver todos os conceptos que queremos que coñezan mediante un tema do seu interese. Evidentemente é imposible atopar algo común para todos, pero

algo que nos últimos anos se repetiu é que moitos alumnos e alumnas están interesados polo fútbol, polo que pensamos que podía ser unha idea. Deseñamos entón unha folla de cálculo con algúns datos dos equipos de fútbol da primeira división: os puntos que tiñan en cada xornada, os partidos gañados, empatados e perdidos, goles a favor e en contra e orzamento. Con estes datos podiamos traballar con estatística unidimensional (por exemplo calculando medias e medianas dos orzamentos) e a estatística bidimensional (vendo por exemplo as relacións entre puntos e goles a favor ou entre goles a favor e orzamento). Ademais, ao incluír os puntos en cada xornada pódese tratar de facer algunha estimación e fantasiar sobre como acabaría a liga se durase catro xornadas máis.

Por suposto, facelo así ten un problema importante, e é que o alumnado ao que non lle gusta o fútbol pode desconectar ou mesmo non entender ben que significan algúns dos parámetros, pero a vantaxe innegable é se consegue relacionar a estatística (algo que moitos só ben como algo que se estuda no instituto) cun tema que resulta atractivo para moitos dos alumnos e alumnas.

3. TRABALLO DE INVESTIGACIÓN

O noso obxectivo nesta materia é que o alumnado sexa capaz de buscar datos, estudalos de xeito rigoroso para extraer conclusións e finalmente presentar o traballo en público diante dos seus compañeiros e mais dos profesores que participan no STEMbach.

O traballo de investigación fanno en grupo e en principio é de tema libre. Esta é habitualmente a parte que resulta máis complicada para o alumnado: pode levar varias sesións que cheguen a escoller un tema sobre o que lles interese traballar. Ás veces damos ideas ou explicamos os temas que traballaron con éxito en cursos anteriores, para que a partir deles fagan as súas propias suxestións.

Entre as propostas tamén tratamos de dar a nosa opinión e facer unha busca previa porque pode ser un tema moi interesante pero sobre que sexa complicado atopar datos. Esa é a seguinte fase: atopar datos fiables. En xeral pensamos que é interesante buscalos en bases de datos públicas (IGE, INE, estudos publicados...), xa que a realización de enquisas (que algún grupo tamén fixo algunha vez) na práctica pode ser difícil e dependendo da mostraxe pode dar lugar a nesgos e conclusións pouco interesantes. Ademais, ao buscar datos na rede tamén aprender a analizar de onde proceden os datos, quen os publicou ou de cando son, co que entenden de que falamos cando dicimos que os datos deben ser fiables. Neste momento de busca pode cambiar o traballo previsto, xa que a partir desta investigación aparecen novas ideas ou ben non atopan os datos que necesitarían para facer a súa investigación inicial.

Unha vez atopados os datos cómpre facer a análise. Habitualmente o que resulta máis interesante e que aprendan a facer regresión sobre os datos para logo estimar como poderían evolucionar no futuro.

Como o traballo se realiza en grupo mentres fan a busca e as primeiras análises vaian creando un documento onde comecen a redactar un borrador do proxecto escrito e no que ademais inclúan as páxinas web das que extraen a información para logo incluílas na bibliografía do traballo.

Desde as primeiras promocións coas que traballamos a análise de datos pensamos que estes traballos resultaban interesantes para presentar á Incubadora de Sondaxes e Experimentos organizado por SGAPEIO, polo que decidimos que o traballo escrito que deben entregar (e debemos avaliar) siga os criterios establecidos no concurso.

4. EXPOSICIÓN DO TRABALLO

A última parte do traballo de investigación proposto é a exposición pública. Unha vez rematado e entregado o traballo escrito dedicamos un tempo á preparación desta exposición, tanto da parte oral como da presentación electrónica.

O alumnado do STEMbach ten que facer un proxecto de investigación en colaboración cunha universidade e a final de 2º de bacharelato ten que expoñelo para que sexa avaliado por un tribunal. Por iso consideramos especialmente importante que o traballo remate coa exposición e establecemos os mesmos criterios que terán que cumprir no último ano: 15 minutos de exposición seguidos de 15 minutos de preguntas do tribunal.

Para a presentación electrónica sempre tratamos de dar indicacións sobre como facer unha boa presentación e como a maior parte do traballo o realizan na clase podemos ir supervisando e comentando con eles como vai avanzando o proceso.

Para parte do alumnado a exposición é o peor momento, xa que non están afeitos a falar en público, e menos tendo en conta que teñen que expoñer o seu traballo no salón de actos do centro, un lugar ben coñecido por eles mais sempre como público.

5. CONCLUSIÓNS

A realización deste traballo de investigación resulta de grande interese para o alumnado. Supón unha primeira aproximación á análise de datos e ao proceso completo de redacción e divulgación dos resultados achados. A pesar de que habitualmente supón un esforzo de imaxinación e de busca de información, o alumnado adoita acabar satisfeito por ver como o profesorado e os seus compañeiros atenden con interese ás súas palabras.

REFERENCIAS

Resolución do 2 de xuño de 2023, da Dirección Xeral de Ordenación e Innovación Educativa, pola que se regula o bacharelato de excelencia en Ciencias e Tecnoloxía (STEMbach), para o curso 2023/24 (consultado o 13/09/2023). Recuperado de https://www.edu.xunta.gal/portal/node/40195

ACTIVIDADES EDUCATIVAS BASEADAS NA ESTATÍSTICA

María Ángel Martínez Rodríguez[1]

[1] IES Agra do Orzán, A Coruña

RESUMO

O Club Matemático do IES Agra do Orzán creouse no ano 2018 como unha sección específica dentro do xa existente, e exitoso, Club de Ciencias, Agracite.

Ao longo dos seis últimos anos os rapaces e rapazas do noso instituto apuntáronse voluntariamente ao club e participamos nun gran número de actividades como o "Día da Ciencia na Rúa" na Coruña, "Matemáticas na Raia" de AGAPEMA, ou a concursos como "Explícoche Matemáticas 2.0" da Universidade de Santiago de Compostela, "La ciencia de los Datos" da Universidade Carlos III ou "Incubadora de Sondaxes e Experimentos" de SGAPEIO.

O alumnado realiza todos estes proxectos con moita ilusión e interese, tanto cando conseguen levar algún recoñecemento como cando non. Gañaron varios premios a nivel galego e incluso nacional, chegando a ser seleccionados para representar a España en Canadá no concurso a nivel mundial de pósters estatísticos do Proxecto Internacional de Alfabetización Estatística ISLP.

Grazas a todas estas actividades e, baseándose na aprendizaxe por descubrimento, desenvolveron novas competencias nas TIC e na exposición e defensa pública de proxectos, ademais de mellorar a súas relacións sociais entre iguais e co mundo, e por suposto na estatística (recollida de datos, tratamento, análise de resultados, representacións gráficas...).

Por todo isto, o Club de Ciencias é un escenario ideal para o desenvolvemento de actividades estatísticas, que supoñen inicialmente un reto tanto para o profesorado como para o alumnado, pero que teñen un grande éxito, cuns resultados persoais e académicos moi salientables.

Mais información:
http://xvicongreso.sgapeio.es/