

*XIV Congreso
Galego
de Estatística
e Investigación
de Operacións*

Libro de actas

Vigo, do 24 ao 26
de outubro de 2019

Escola de Enxeñería Industrial
Sede cidade

XIV Congreso Galego de Estatística e Investigación de Operacións

Libro de actas

Vigo, 24–26 de outubro de 2019

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

Elaborado por:
Comité Organizador

I.S.B.N.: 978-84-09-15581-1

Comité científico

- JACOBO DE UÑA ÁLVAREZ (Presidente. Universidade de Vigo)
- GUSTAVO BERGANTIÑOS CID (Universidade de Vigo)
- CARMEN CADARSO SUÁREZ (Universidade de Santiago de Compostela)
- RICARDO CAO ABAD (Universidade da Coruña)
- IGNACIO GARCÍA JURADO (Universidade da Coruña)
- WENCESLAO GONZÁLEZ MANTEIGA (Univ. de Santiago de Compostela)
- MARÍA ESTHER LÓPEZ VIZCAÍNO (Instituto Galego de Estatística)
- RAQUEL MENEZES (Universidade do Minho)
- SALVADOR NAYA FERNÁNDEZ (Universidade da Coruña)
- COVADONGA RODRÍGUEZ-MOLDES REY (IES de Mugardos)
- CÉSAR SÁNCHEZ SELLERO (Universidade de Santiago de Compostela)
- GIOVANI SILVA (Universidade de Lisboa)

Comité organizador

- JAVIER ROCA PARDIÑAS (Presidente)
- ANA MARÍA DOMÍNGUEZ PUENTE
- MARÍA JOSÉ GINZO VILLAMAYOR
- MARÍA GÓMEZ RÚA
- MARÍA DEL CARMEN IGLESIAS PÉREZ
- MARÍA JESÚS LONGA SAYÁNS
- LETICIA LORENZO PICADO
- JUAN CARLOS PARDO FERNÁNDEZ
- ALEJANDRO SAAVEDRA NIEVES
- ANTONIO VAAMONDE LISTE

Presentación

Por petición da Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO), o Departamento de Estatística e Investigación Operativa da Universidade de Vigo organiza o *XIV Congreso Galego de Estatística e Investigación de Operacións*, que se celebra na Sede Cidade da Escola de Enxeñería Industrial de Vigo entre os días 24 e 26 de outubro de 2019.

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

O Comité Científico
O Comité Organizador

Contidos

Conferencias plenarias	8
<i>A estatística oficial nunha sociedade datificada.</i> Juan Manuel Rodríguez Poo	9
<i>Designing routes for vehicles and drivers.</i> Juan José Salazar González	9
<i>Popurrí estatístico e bayesiano con harmonías lonxitudinais e de supervivencia a ritmo de ciencia de datos.</i> Carmen Armero Cervera	10
<i>On semiparametric modelling, estimation and inference for survival data subject to dependent censoring.</i> Ingrid Van Keilegom	10
Sesión especial de Biometría	11
<i>“E cara onde imos?”, díxolle unha pulga a outra.</i> Rosa Crujeiras	12
<i>Applying state space models to monthly temperature data in European cities.</i> Magda Monteiro	14
<i>Geostatistical analysis of sardine eggs data – a Bayesian approach.</i> Soraia Pereira	18
Comunicacións orais	19
<i>Un nuevo algoritmo heurístico para el truck and trailer routing problem.</i> Laura Davila Pena, Balbina Casas Méndez, David Rodríguez Penas	20
<i>Una heurística para la planificación óptima de tareas en una empresa de ayuda a domicilio.</i> Isabel Méndez-Fernández, Silvia Lorenzo Freire, Ignacio García Jurado	21
<i>Optimización web dun sistema de xestión de stocks.</i> Luís Caviqne, José Saias, Jorge Santos, Armando B. Mendes	24
<i>Problema de asignación de muestras de laboratorio.</i> Diego Noceda Davila, Luisa Carpena, Silvia Lorenzo Freire	27
<i>Cooperative approach to a plant location problem with agglomerative economies.</i> Adriana Navarro Ramos, Gustavo Bergantiños	29
<i>Uncertainty in cooperative interval games: How Hurwicz criterion compatibility leads to egalitarianism.</i> Lina Mallozzi, Juan J. Vidal Puga	33
<i>Os tabuladores, as súas compoñentes e mais a súa utilidade.</i> Carlos L. Iglesias Patiño	47

<i>O inventario de operacións e actividades estatísticas. Indicadores de calidade.</i>	
María Martín Vila, Estefanía Villar Cheda, Carlos L. Iglesias Patiño	53
<i>Proxeccións de taxas de actividade.</i> Pablo Torrado Solar, Eva Alberte Aparicio	63
<i>Análise e modelaxe de opinións sobre a calidade dos servizos prestados polo sector hosteleiro ourensán.</i> Diego Frade Amil, Tomás R. Cotos-Yáñez, Manuel Mosquera-Rodríguez, Ana Pérez-González	73
<i>Unha aproximación á avaliación do impacto recente de fondos europeos en Galicia.</i> N. Veiguela Fernández, M. E. López Vizcaíno, I. del Río Viqueira, C.L. Iglesias Patiño, R. Jácome Rodríguez, M.E. Calvo Ocampo, A. Barros Torres, J.M. Martínez Barral, J. Landeira Caamaño, E. Andión Hermida, A. Tarrío Tobar	76
<i>Modelización da degradación de biomateriais usados en rexeneración ósea.</i> Salvador Naya Fernández, Javier Tarrío Saavedra, Yaroslava Robles, Silvia Díaz Prado	93
<i>Analysis of extreme temperature events on Greenland ice sheet.</i> María Isabel Borrajo, Emma Eastoe, Amber Leeson	99
<i>Variabilidade onomástica VS variabilidade lingüística: unha visión espacial.</i> M.J. Ginzo Villamayor, R.M. Crujeiras, X. Sousa Fernández	102
<i>Extent of occurrence estimation for invasive plants.</i> Paula Saavedra Nieves, Alberto Rodríguez-Casal	104
<i>Correction when using estimates in logistic regression for longitudinal data.</i> Rosa C. S. Oliveira	105
<i>R packages to missing longitudinal data.</i> Susana R. G. Martins	112
<i>Nonparametric estimation of the cure rate when cure status is partially known.</i> Wende Clarence Safari, Ignacio López-de-Ullibarri, M. Amalia Jácome Pumar	113
<i>A functional diagnostic test based on ROC curves analysis.</i> Graciela Estévez Pérez, Philippe Vieu	123
<i>La calidad de la democracia: ¿es el género determinante para medir la calidad de la democracia en Perú?</i> Luz Anamar González de la Rosa, Luis Javier Sánchez-Martínez	129
<i>RAPOSa, una herramienta gratuita para resolver problemas de optimización polinómica.</i> Brais González Rodríguez, Julio González Díaz, Angel Manuel González Rueda, Joaquín Ossorio Castillo, Beatriz Pateiro López, Diego Rodríguez Martínez, David Rodríguez Penas	138
<i>Estimating the random arrival rule in bankruptcy: a case study.</i> Alejandro Saavedra Nieves	139

<i>Metodoloxías bioinformáticas y bioestadísticas en el análisis del inbreeding genómico.</i> L.J. Sánchez Martínez, C.L. Hernández, G. Pita, R. Calderón .	141
<i>Small area estimation under bivariate nested error regression models.</i> María Dolores Esteban, María José Lombardía, Esther López Vizcaíno, Domingo Morales, Agustín Pérez	143
<i>Un novo contraste de bondade de axuste para modelos de regresión cuantil cando a variable resposta está censurada pola dereita.</i> Mercedes Conde Amboage, Ingrid Van Keilegom, Wenceslao González Manteiga	145
<i>Sparse semiparametric regression for functional data.</i> Silvia Novo Díaz, Germán Aneiros, Philippe Vieu	147
<i>Backfitting tests in generalised structured models.</i> Enno Mammen, Stefan Sperlich	150
<i>Improved q-values for discrete uniform and homogeneous tests.</i> Marta Cousido Rocha, Jacobo de Uña Álvarez, Sebastian Döhler	153
<i>Bandwidth selection for nonparametric kernel prediction.</i> Inés Barbeito Cal, Ricardo Cao Abad, Stephan Sperlich	154
<i>A test for comparing conditional ROC curves with multidimensional covariates.</i> Arís Fanjul Hevia, Wenceslao González-Manteiga, Juan Carlos Pardo-Fernández, Ingrid Van Keilegom	157
<i>Nonparametric estimation of the probability of default in credit risk.</i> Rebeca Peláez Suárez, Ricardo Cao Abad, Juan M. Vilar Fernández	159
<i>Estimación no paramétrica de máxima verosimilitud de la supervivencia libre de enfermedad con datos de corte transversal.</i> Beatriz Piñeiro Lamas, Jacobo de Uña Álvarez	163

Pósters 175

<i>Estatística en la planificación estratégica de ABANCA.</i> Belén Fernández de Castro, Teresa Veiga Rodríguez, Alejandro Figueroa Silva, Sandra Amado Souto	176
<i>Estudo da diversidade xenética en dúas poboacións de Zostera noltei.</i> Iago Álvarez Lomba, Juan J. Hermida, Edita López	179
<i>Evidencias locais do cambio climático a partir do rexistro de datos das estacións meteorolóxicas do Rosal e de Oia.</i> Mario Fernández Baz, Edita López, Juan J. Hermida, Ángeles Saavedra	182
<i>Índices de Sobol na análise de influência das propiedades em estruturas compostas.</i> Luísa N. Hoffbauer, Carlos C. António	186
<i>Covariate measurement error in a nonlinear model for longitudinal data.</i> Rosa Santos Oliveira	192

<i>Unha análise non paramétrica sobre a eficiencia e a produtividade dos países europeos en tecnoloxía e propiedade intelectual.</i> Agostinho Rosa, Armando B. Mendes, Jorge Santos, Luis Cavique	198
<i>Nuevos modelos derivados del truck and trailer routing problem.</i> Laura Davila Pena, Balbina V. Casas Méndez	202
<i>Estimación del core-center en juegos de aeropuerto.</i> Nadine Helena Espinoza Burgos, Estela Sánchez Rodríguez	203
<i>Algoritmos para el cálculo del core-center de un juego de bancarrota.</i> Iago Núñez Lugilde, Estela Sánchez Rodríguez	204
<i>Optimización de rutas de carros mezcladores en una cooperativa agrícola.</i> Carlos Amiama, Luisa Carpenté, Silvia Lorenzo Freire	206
<i>Mejorando la motivación de tus estudiantes con técnicas de innovación educativa.</i> Marta Cousido Rocha, María Gómez Rúa	208
<i>Plugin de R-Commander para la enseñanza de estadística básica: RcmdrPlugin.TeachStat.</i> T.R. Cotos Yáñez, M.A. Mosquera Rodríguez, A. Pérez González, B. Reguengo Lareo	211
<i>¿Avanzan los países hacia el logro de los objetivos establecidos en políticas ambientales? Análisis multivariante de una década.</i> M.J. Fernández-Gómez, J. Martín-Vallejo, I. Barrera-Mellado, I. Gallego-Álvarez, B. Sastre-Hernández	212
<i>A generalized product-limit estimator of the survival function when cures are partially known.</i> Wende Clarence Safari, M. Amalia Jácome Pumar, Ignacio López-de-Ullibarri	213
<i>Rexións de incidencia dos incendios forestais en Galicia.</i> Manuel Antonio Novo Pérez, Ana Belén Buide Carballosa, María José Ginzo Villamayor	224
<i>Intervalos de confianza bootstrap para curvas percentís do IMC na poboación infantil galega.</i> María del Carmen Iglesias Pérez, María Arias Vázquez, María Isolina Santiago Pérez	226
<i>Una revisión de los contrastes de equivalencia de las medias de dos poblaciones.</i> Javier Rey Ramírez, Juan Carlos Pardo-Fernández	234
<i>Efecto de la suavización en contrastes basados en los residuos de la regresión.</i> Natalia Pérez Veiga, Juan Carlos Pardo-Fernández	235
<i>Hyperscale 3D point classification through functional data analysis.</i> Manuel Oviedo, Javier Roca, Carlos Cabo, Celestino Ordóñez	240

Conferencias plenarias

A estatística oficial nunha sociedade datificada

Juan Manuel Rodríguez Poo (Instituto Nacional de Estadística)

Unha sociedade datificada ofrece novas oportunidades de desenvolvemento na elaboración de estatísticas por parte dos Institutos Nacionais de Estatística. Consciente desta situación, a estadística oficial está presentando novos pasos orientados a manter a relevancia da produción estatística e, así, continuar satisfacendo as fortes demandas de información da máis alta calidade por parte dos usuarios dos datos.

Designing routes for vehicles and drivers

Juan José Salazar González (Universidad de La Laguna)

Most of the literature concerns the design of routes through customers for one resource (for example, a fleet of capacitated vehicles). However, it is quite common in real-world applications the need of optimizing more than one resource (for example, vehicles and drivers), both very interrelated but each one with different requirements. This talk will address some optimization problems in this category. The motivation of our research started when solving an airline problem in Canary Islands operating about 150 flights (customers) every day. Each flight must be operated by one aircraft and one crew, and the problem is to compute optimal assignments (routes). There are eight airports in the islands, and two of them are hubs (Tenerife and Gran Canaria). To reduce overnight costs, a crew route must start and end in the same hub airport. Due to maintenance operations, each aircraft route must start and end in different hub airports.

We will present different mathematical models and algorithms to find optimal routes detailed in *Omega* 43 (2014) 71–82 and *Transportation Science* 51 (2017) 250–268. In the above problem, the flights are assumed to have a fixed departure time. We will also analyze the variant where these times can be slightly changed to find better solutions. This variant motivates a new problem where one must determine the time to start serving each customer, and design optimal routes for vehicles and drivers. In this problem synchronization constraints are crucial to guarantee feasibility of the solution. Note that in all these problems, each customer (flight) is visited (operated) by one vehicle and one driver. Although they are formulated and solved using standard elements in vehicle routing, they are scheduling problems in the sense that each customer is a task to be performed (for example, going from one airport to another airport). The new problem is called “Vehicle-and-Driver Scheduling Problem” and our research results are detailed in *Computers & Operations Research* 92 (2018) 56–64. We will also introduce another new problem which is a more natural routing problem, in the sense that each customer is a location in the space, and some locations represent exchange locations where drivers

exchange vehicles. The new problem is called “Driver-and-Vehicle Routing Problem” and our research results are detailed in *Computers & Operations Research* 81 (2017) 247-256.

Popurrí estatístico e bayesiano con harmonías lonxitudinais e de supervivencia a ritmo de ciencia de datos

Carmen Armero Cervera (Universitat de València)

Esta comunicación é, como o seu nome indica, un popurrí de varias cousas. Comezaremos por presentar o Grao de Ciencia de Datos da Universitat de València, estrutura e materias de Matemáticas e Estatística, en particular a inferencia bayesiana. Este xeito de entender a estatística, cada vez máis popular e útil, baséase nunha definición de probabilidade que pode asignar probabilidades e distribucións de probabilidade a calquera elemento estatístico con incerteza (sucesos e variables aleatorias, parámetros, hipóteses, observacións faltantes etc.) e é capaz de incorporar nova información ao proceso inferencial mediante o uso secuencial do teorema de Bayes. A estatística bayesiana ofrece boas ferramentas para o tratamento dos modelos conxuntos de datos lonxitudinais e de supervivencia. Son modelos deseñados para estudar fenómenos temporais que inclúen características xerais da poboación obxecto de estudo e tamén características específicas individuais, xerando, deste xeito, inferencias e predicións tanto poboacionais coma individuais.

On semiparametric modelling, estimation and inference for survival data subject to dependent censoring.

Ingrid Van Keilegom (KU Leuven)

When modelling survival data, it is common to assume that the survival time T is conditionally independent of the censoring time C given a set of covariates. However, there are numerous situations in which this assumption is not realistic. The goal of this paper is therefore to develop a flexible semiparametric transformation model, which states that after a proper nonparametric monotone transformation, the vector (T, C) follows a linear model, and the vector of errors in this bivariate linear model follows a bivariate normal distribution with non-diagonal variance-covariance matrix. We show that this semiparametric model is identified, and propose estimators of the nonparametric transformation, the regression coefficients and the correlation between the error terms. The performance of the proposed method is investigated both in an asymptotic way and through finite sample simulations.

Sesión especial de Biometría

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

”E CARA ONDE IMOS?” DÍXOLLE UNHA PULGA A OUTRA

Rosa M. Crujeiras¹, María Alonso-Pena¹ e Jose Ameijeiras-Alonso²

¹Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

²Department of Mathematics, KU Leuven (Bélxica)

RESUMO

En 1981, Edward Batschelet recolleu nunha monografía un resumo de diversas técnicas estatísticas para a análise de datos circulares. Os exemplos motivadores abrangían problemas de análise descritiva e de modelización de ritmos circadianos, fenoloxía, orientación de plantas e, sobre todo, orientación animal. En todos estes contextos, os datos recollidos poden ser representados como puntos nun círculo de radio unidade, e polo tanto, analizados coas metodoloxías propias dos datos circulares.

Cun enfoque fundamentalmente paramétrico e centrándose no uso do modelo de densidade circular von Mises, Batschelet (1981) presenta solucións aos problemas de identificación de orientacións preferentes, a través da estimación dos parámetros da von Mises, previo contraste da idoneidade desta distribución. En volumes posteriores, Fisher (1995), Jammalamadaka e Sengupta (2001) ou máis recentemente, Pewsey et al. (2013), tamén consideran os exemplos sobre orientación de animais para propoñer outro tipo de técnicas de modelado, tanto en densidade como en regresión.

Con todo, as importantes achegas de Batschelet xa foron obxecto de crítica pouco despois da publicación do seu libro. Oden (1983) sinala que o uso da distribución von Mises pode ser cuestionable, dado que non existe un resultado análogo ao Teorema Central do Límite para datos reais, e que xustifica o uso da normal neste contexto.

Nesta contribución, revisaremos as ferramentas de tipo non paramétrico das que dispoñemos actualmente para a modelización da densidade e da regresión con datos circulares. Nos exemplos de orientación animal, o obxectivo soe ser determinar cal é a dirección preferente (cara onde van os animais baixo determinadas circunstancias), un problema para o que se poden considerar metodoloxías ligadas á identificación de modas (véxase Ameijeiras-Alonso et al. (2018), que propón contrastes sobre multimodalidade para datos circulares). Se esta dirección pode vir influenciada por outros factores, e interesa coñecer o impacto dos mesmos, a alternativa metodolóxica son os modelos de regresión (véxase Di Marzio et al. (2013) onde se presenta unha estimación non paramétrica da regresión con resposta circular).

Estas dúas perspectivas serán revisadas e utilizaranse para determinar a dirección de escape de dúas variedades de pulgas de praia, estudadas por Scapini et al. (2002), en dous experimentos levados a cabo en primavera e outono, baixo distintas condicións de temperatura e humidade, e utilizando algunhas barreiras visuais para os insectos.

Palabras e frases chave: Orientación animal; Datos circulares; Estatística nonparamétrica.

REFERENCIAS

- Ameijeiras-Alonso, J., Benali, A., Crujeiras, R.M., Rodríguez-Casal, A. e Pereira, J.M.C. (2018) Fire seasonality identification by multimodality tests. arXiv:1810.04498
- Batschelet, E. (1981) Circular Statistics in Biology. Academic press.
- Di Marzio, M., Panzera, A. e Taylor, C.C. (2013) Non-parametric regression for circular responses. Scandinavian Journal of Statistics, 40, 238-255.
- Fisher, N. I. (1995) Statistical Analysis of Circular Data. Cambridge University Press.
- Jammalamadaka, S.R. e Sengupta, A. (2001) Topics in Circular Statistics. World scientific.
- Oden, N. (1983) Review on: Circular Statistics for Biology. The Quaterly Review of Biology, 58(2), 312.
- Pewsey, A., Neuhäuser, M. e Ruxton, G. D. (2013). Circular Statistics in R. Oxford University Press.
- Scapini, F., Aloia, A., Bouslama, M.F., Chelazzi, L., Colombini, I., El Gtari, M., Fallaci, M. e Marchetti, GM (2002) Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, *Talitrus saltator* and *Talorchestia bito*, from an exposed Mediterranean beach. Behavioural Ecology and Sociobiology, 51(1), 403-414.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

CHANGE POINT ANALYSIS IN A STATE SPACE FRAMEWORK TO MONTHLY TEMPERATURE DATA IN EUROPEAN CITIES

Magda Monteiro^{1,2}, Marco costa^{1,2}

¹ ESTGA - Águeda School of Technology and Management, University of Aveiro

² CIDMA - Center for Research & Development in Mathematics and Applications, University of Aveiro

Abstract

In this work, we present time series of monthly average temperatures in several European locations which were statistically analyzed using a state space approach, where it is considered a model with a deterministic seasonal component and a stochastic trend. The analysis of smoother prediction of the stochastic trend and its comparison in a temporal viewpoint can reveal patterns about warming in Europe. The temperature rise rates in Europe seem to have increased in the last decades when compared with longer periods, hence a change point detection method is applied to the trend component in order to identify these possible changes in the monthly temperature rise rates. The adopted methodology pointed out, for most series a change point in the late eighties.

Keywords : climate change, monthly data, temperature data, time series analysis, state space modeling, Change point detection

1. INTRODUCTION

Global warming and the associated climate changes are topics in the world's agenda, which are receiving much attention from the scientific community, due to their major social, economic and health impacts on human life.

These changes are already being felt all over the world, with the increase of natural catastrophes such as droughts, floods, tropical storms, among others. Nevertheless, global warming effects vary around the world, therefore this phenomenon must be monitored at a smaller scale. In particular, given our location, we are interested in understanding these phenomena at Europe scale. In this sense, the analysis of local temperature time series is an important topic that can contribute to a better understanding of climate dynamics at smaller scales, allowing for the identification of their change patterns. Thus, this analysis can be very useful for the application of efficient environment monitoring processes.

This work is a preliminary analysis of long monthly average temperature time series in fifteen European cities in order to find change point associated with rise rates changes of temperature. The Data is available at the Climate Data Online and comprises the period between January 1900 and December 2017, making a total of 1416 observations. The location of the cities and data descriptive statistics can be found in Table 1. The major data of this work has already been analyzed in Costa and Monteiro(2017), where state space models associated to the Kalman filter algorithm and also a cluster procedure to smoother trend levels were applied in order to investigate patterns on the temperature rise in Europe. A conclusion therein is that temperature rise rates in Europe seem to have increased in the last decades being important to find the change points and use the information from there to get accurate temperature rise rates. Therefore, in the present work it is applied a change point detection procedure of maximum type in order to assess if there is a change point in the temperature trend pattern for each city under analysis.

There are several works on change point detection in climate or environmental settings. For instance, Reeves et al.(2007) makes a review of this topic for models with independent and identically distributed (i.i.d.) errors and Jarušková and Antoch(2018) applied several change points

procedures to Klementinum temperature. In line with this last work, it will be considered basic statistical tests with the applications of maximum type statistics to the stochastic trend of the temperature time series in Europe.

city (country)	coord.	min	max	average	average	missing
		1900		1900	2000	values
		to 2017		to 1917	to 2017	(%)
Berlin (Germany)	52.47N, 13.30E	-10.7	23.0	8.7	9.9	2.3
Bucharest (Romania)	44.40N, 26.10E	-11.5	27.5	10.1	11.1	2.6
Copenhagen (Denmark)	55.68N, 12.55E	-6.7	21.6	8.6	9.6	0.4
Dublin (Ireland)	53.43N, 6.25W	0.0	17.1	9.2	9.6	2.2
Kiev (Ukraine)	50.40N, 30.57E	-15.3	26.0	8.1	9.4	0.9
Lisbon (Portugal)	38.72N, 9.15W	8.0	25.1	16.9	17.4	3.4
Minsk (Belarus)	53.93N, 27.63E	-16.8	22.6	6.1	7.4	1.8
Nantes (France)	47.17N, 1.60W	-1.2	23.3	12.0	12.5	0.4
Prague (Czech Rep.)	50.10N, 14.25E	-13.1	22.0	7.9	9.1	8.2
S. Petersburg (Russia)	59.97N, 30.30E	-18.3	24.4	5.3	6.5	0.0
Talin (Estonia)	59.42N, 24.80E	-15.5	21.5	5.3	6.5	5.4
Vienna (Austria)	48.25N, 16.37E	-9.5	24.4	10.1	11.3	0.2
Vilnius (Lithuania)	54.63N, 25.10E	-17.1	21.8	5.8	7.3	2.6
Zagreb (Croatia)	45.82N, 15.98E	-7.0	25.9	11.8	13.0	1.4
Zurich (Switzerland)	47.38N, 8.57E	-8.7	22.4	8.7	9.8	0.0

Table 1: Characterization of time series data set.

2. MODEL DESCRIPTION

In order to model, in each city, the monthly temperature time series, it is considered a model with a deterministic seasonal component and a stochastic trend, defined by

$$Y_t = \tau_t + \sum_{i=1}^{12} \beta_i S_{t,i} + e_{1,t} \quad (1)$$

$$\tau_t = \mu + \phi(\tau_{t-1} - \mu) + e_{2,t} \quad (2)$$

where $\{\tau_t\}$ is the trend component that follows a stationary autoregressive process with mean μ and the autoregressive parameter $|\phi| < 1$; the error $e_{i,t}$, with $i = 1, 2$, follows a white noise process ($E(e_{i,t}) = 0$, $var(e_{i,t}^2) = \sigma_i^2$ and $E(e_{i,t}e_{i,r}) = 0$, for $t \neq r$), and are uncorrelated errors, that is $E(e_{1,t}e_{2,r}) = 0$, $\forall t, r$. The error $e_{1,t}$ is called the observation error and it can be seen as a measure error, whereas the error $e_{2,t}$ is called the state error and translates the randomness of the trend component.

Note that state τ_t is an unobservable variable and its predictions must be obtained. For that, the state process $\{\tau_t\}$ in the model (1)-(2) can be predicted by Kalman filter and Kalman smoother estimators.

This model was also used in Costa and Monteiro(2017) to which a clustering procedure was applied to understand the patterns associated to the climate in Europe. In terms of parameter estimation, similar to what was done in this previous work, a classical decomposition approach was used combining the least squares estimation of the seasonal parameters β_i , with $i = 1, \dots, 12$, with a distribution-free estimators developed to state space models in order to estimate the remaining parameters.

2. CHANGE POINT DETECTION

In order to discover an abrupt change in the behaviour of the temperature rise rates, a basic statistical test applying maximum type statistics to detect a change in location is performed. If

there are no structural changes in the temperature trend, it is expected that τ_t has no changes over time in its mean. Therefore, it is applied a statistical test to the smoother predictions, $\tau_{t|n}$, of τ_t having into account that $\tau_1, \tau_2, \dots, \tau_n$ are an AR(1) process and that smoother predictions are the predictors of the non-observable component that present lower mean square error when compared with the other Kalman filter predictions.

Briefly, if we are in the presence of an AR(1) process, $\{X_t\}$, the hypotheses of the test of *maximum type* are

$$H_0: X_t \text{ are an AR(1) sequence normally distributed variables according to the same } N(\mu, \sigma^2)$$

vs

$$H_1: \text{ there is a time point } k \in \{1, \dots, n-1\} \text{ such that } X_1, \dots, X_k \text{ are distributed according to } N(\mu_1, \sigma^2) \text{ and } X_{k+1}, \dots, X_n \text{ are distributed according to } N(\mu_2, \sigma^2).$$

Assuming σ^2 is unknown, then the test statistic $T(n)$ is the maximum of the absolute values of two sample t -test statistics

$$T(n) = \max_{1 \leq k < n} |T_k| = \max_{1 \leq k < n} \sqrt{\frac{(n-k)k}{n}} \frac{|\bar{X}_k - \bar{X}_k^*|}{s_k}$$

where

$$\bar{X}_k = \frac{1}{k} \sum_{i=1}^k X_i, \bar{X}_k^* = \frac{1}{n-k} \sum_{i=k+1}^n X_i \text{ and } s_k = \sqrt{\frac{1}{n-2} \left[\sum_{i=1}^k (X_i - \bar{X}_k)^2 + \sum_{i=k+1}^n (X_i - \bar{X}_k^*)^2 \right]}.$$

The null hypothesis can be rejected if the statistic $T(n)$ is greater than the critical value. The exact distribution of $T(n)$, for independent variables, is very complex and Worsley(1979) was able to calculate the true critical values only for the number of observations n less than 10. Alternatively, approximate critical values can be computed by other methods as the Bonferroni inequality, simulation or the asymptotic distribution.

Antoch et al.(1997) shows that when random variables $X_1, X_2, \dots, X_n$ are not independent but form an ARMA sequence then the asymptotic critical values of the test statistics considering independence have to be multiplied by $\sqrt{2\pi f(0)/\gamma}$, where $\gamma = \text{var} X_t$ and $f(\cdot)$ denotes the spectral density function of the corresponding ARMA process. Especially for an AR(1) sequence, the critical values should be multiplied by $[(1+\phi)(1-\phi)^{-1}]^{1/2}$ where ϕ is the first autoregressive coefficient, (Jarušková,1997).

The present work uses critical values obtained by the asymptotic distribution which expression can be found in Jarušková (1997) and multiplied by the AR(1) factor with the estimates of ϕ .

For each city under study, these basic change point procedure detected a change and Table 2 summarizes these changes and also present the mean of smoother predictors before and after changes. The majority of temperature time series presents a change point in the late eighties. Lisbon presents a change point much earlier, in the previous decade, and Bucharest presents a change point about twenty years later.

Note that western cities present the lowest mean values of $\tau_{t|n}$ and because of that the shorter distances between averages before and after the detected change point with differences below 0.5 °C. In the opposite site are the major eastern cities with differences above 1 °C.

This work is at an early stage and there are some issues that need further research. For instance we are applying the change point test not directly to the stochastic trend component, that is unobservable, but to the smoother predictions. The way the Kalman smoother predictions are built induces a strong dependence between consecutive values and their use to evaluate the test assumptions associated with the trend component might be misleading, so it is necessary to study this subject in depth. Another issue to be studied is the possibility of other change points, since it was only analyzed the at most one change point case. Finally, the confrontation between the obtained results and other different change points methodologies will be the target of future work.

city (country)	change point month/year	mean bef. c.-p.	mean after c.-p.
Berlin (Germany)	9/1987	0.491	1.426
Bucharest (Romania)	9/2006	0.517	1.732
Copenhagen (Denmark)	9/1987	0.416	1.376
Dublin (Ireland)	10/1988	0.281	0.721
Kiev (Ukraine)	12/1989	0.506	1.479
Lisbon (Portugal)	7/1978	0.357	0.843
Minsk (Belarus)	11/1987	0.659	1.814
Nantes (France)	7/1987	0.359	0.827
Prague (Czech Rep.)	8/1987	0.563	1.457
S. Petersburg (Russia)	12/1988	0.572	1.643
Talin (Estonia)	12/1988	0.507	1.596
Vienna (Austria)	8/1987	0.520	1.297
Vilnius (Lithuania)	11/1987	0.715	2.325
Zagreb (Croatia)	12/1992	0.389	1.274
Zurich (Switzerland)	8/1987	0.578	1.329

Table 2: Summary of change point detection.

ACKNOWLEDGEMENTS

Authors were partially supported through Portuguese funds by the Center for Research and Development in Mathematics and Applications (CIDMA) and the Portuguese Foundation for Science and Technology (“Fundação para a Ciência e a Tecnologia” [FCT]), within project UID/MAT/04106/2019.

REFERENCES

- Alpuim, T. and El-Shaarawi, A. (2009). Modeling monthly temperature data in Lisbon and Prague. *Environmetrics*, **20**: 835–852.
- Antoch, J., Hušková, M. and Z. Prášková (1997). Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*. **60**, 291–310.
- Costa, M. and Monteiro, M. (2016). Bias-correction of Kalman filter estimators associated to a linear state space model with estimated parameters. *Journal of Statistical Planning and Inference* **176**: 22 – 32.
- Costa, M. and Monteiro, M. (2017). Statistical Modeling of an Air Temperature Time Series of European Cities, In *Advances in Environmental Research*. vol. 59, ed. Justin A. Daniels, 213 – 236. ISBN: 978-1-53612-701-0. New York: Nova Science Publishers.
- Costa M. and Monteiro M. (2018). A periodic mixed linear state-space model to monthly long-term temperature data. *Environmetrics* e2550 <https://doi.org/10.1002/env.2550>
- Jarušková, D. (1997). Some problems with application of change point detection methods to environmental data. *Environmetrics*, 8:469–483.
- Jarušková, D. and Antoch, J. (2018). Changepoint analysis of Klementinum temperature series. *Environmetrics*. 2019:e2570. <https://doi.org/10.1002/env.2570>.
- Reeves, J. Chen, J., Wang, X.L., Lund, R.B. and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*. 46, 900–915.
- Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*. **74**: 365–367.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

**GEOSTATISTICAL ANALYSIS OF SARDINE EGGS DATA.
A BAYESIAN APPROACH**

Soraia Pereira¹

¹FCUL, Universidade de Lisboa

ABSTRACT

Understanding the distribution of animals over space, as well as how that distribution is influenced by environmental covariates, is a fundamental requirement for the effective management of animal populations. This is especially the case for populations which are harvested. The sardine is one of the most important fisheries species, both for its economic, sociologic, anthropologic and cultural values.

Here we intend to understand the spatial distribution of the average number of sardine eggs by m^3 . Our main objectives are to identify the environmental variables that better explain the spatial variation in sardine eggs density and to make predictions in spatial points that were not observed.

The data structure presents an excess of zeros and extreme values. To deal with this, we propose a point-referenced zero-inflated model to model the probability of presence together with the positive sardine eggs density and a point-referenced generalized Pareto model for the extremes. Finally, we combine the results of these two models to get the spatial predictions of the variable of interest. We follow a Bayesian approach and the inference is made using the package R-INLA in the software R.

Comunicacións orais

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

UN NUEVO ALGORITMO HEURÍSTICO PARA EL TRUCK AND TRAILER ROUTING PROBLEM

Laura Davila Pena¹, Balbina V. Casas Méndez¹ y David Rodríguez Penas¹

¹IMAT, Instituto de Matemáticas, MODESTYA, Grupo de Modelos de Optimización, Decisión, Estadística y Aplicaciones, Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela.

RESUMEN

Los problemas de rutas de vehículos, VRPs (Dantzig y Ramser, 1959), admiten diversas variantes, según las necesidades de los clientes y otras características que se pueden presentar al analizar situaciones de la vida real. Una de ellas es el denominado truck and trailer routing problem, TTRP (Chao, 2002), donde una flota de camiones y tráileres atienden a un conjunto de clientes tal que cuando un tráiler no puede acceder a un cliente, este es atendido solo por un camión.

Motivado por la problemática existente en el ámbito de la logística, este trabajo propone nuevos modelos y algoritmos de optimización para problemas TTRP. En cuanto a la modelización, se está desarrollando el caso que contempla la división en compartimentos de los vehículos (Coelho y Laporte, 2015). Por otro lado, la necesidad de resolver TTRPs que involucren un número relativamente alto de clientes ha servido de motivación para recurrir a los métodos heurísticos.

En particular, se presenta en esta comunicación una adaptación del algoritmo de Clarke & Wright (Clarke y Wright, 1964) para el caso del TTRP, por ser este método eficaz a la hora de construir soluciones factibles de un VRP con restricciones de capacidad. Esta nueva heurística, llamada CW_TTRP, se ha programado haciendo uso del lenguaje R, y ha sido comparada con otros procedimientos de solución del estado del arte, obteniendo resultados satisfactorios tanto en calidad de la solución, como en el esfuerzo computacional necesario para obtenerla. Además, en el proceso de validación se han utilizado diversos ejemplos, entre los que se encuentra un problema de transporte del mundo real.

Palabras e frases chave: problemas de transporte, modelos de rutas de vehículos, optimización lineal entera mixta, algoritmo heurístico.

Agradecimientos: La investigación de Laura Davila Pena está financiada por el Ministerio de Educación, Cultura y Deporte del Gobierno de España (contrato FPU17/02126). Los autores agradecen la financiación recibida del FEDER (MINECO/AEI proyecto MTM2017-87197-C3-3-P) y de la Xunta de Galicia (Grupo Competitivo ED431C 2017/38).

REFERENCIAS

- Chao, I. M. (2002) A tabu search method for the truck and trailer routing problem. *Computers & Operations Research*, 29, 33–51.
- Clarke, G. and Wright, J. W. (1964) Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12, 568–581.
- Coelho, L. C. and Laporte, G. (2015) Classification, models and exact algorithms for multicompartment delivery problems. *European Journal of Operational Research*, 242, 854–864.
- Dantzig G. B. and Ramser, J. H. (1959) The truck dispatching problem. *Management Science*, 6, 80–91.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

UNA HEURÍSTICA PARA LA PLANIFICACIÓN ÓPTIMA DE TAREAS EN UNA EMPRESA DE AYUDA A DOMICILIO

Isabel Méndez Fernández, Silvia Lorenzo Freire e Ignacio García Jurado

Grupo MODES, Departamento de Matemáticas, Facultad de Informática, Universidade da Coruña.

RESUMEN

En este trabajo se presenta un problema de planificación de horarios para una empresa que presta servicios de atención a domicilio. Se quiere determinar el número de auxiliares que se debe contratar y su planificación, a la vez que se consideran varios objetivos a optimizar. Para resolver el problema se diseña un modelo de programación lineal entera, que solamente puede aplicarse a problemas de pequeño tamaño, y un algoritmo heurístico, para resolver problemas de tamaño real.

Palabras y frases clave: Problema de Planificación, Técnicas Heurísticas, Simulated Annealing, Búsqueda Local

1. INTRODUCCIÓN

Los servicios de atención a domicilio tienen como objetivo ayudar a personas mayores o dependientes a mantener su calidad de vida sin necesidad de abandonar sus hogares. Puesto que las personas suelen sentirse mas cómodas si pueden continuar viviendo en sus casas, en lugar de tener que trasladarse a un centro especializado.

En este trabajo nos centramos en el problema que presenta Mayores, una empresa que presta servicios de atención a domicilio en A Coruña. Para proporcionar sus servicios la empresa dispone de unas auxiliares, que se encargan de visitar a los usuarios en sus domicilios y realizar las tareas que les han sido asignadas (como preparar comidas, supervisar medicación, tareas de limpieza, etc.). Los usuarios de la empresa requieren de la realización de un conjunto de servicios, ocasiones en las cuales las auxiliares deben atenderlos en sus hogares, y, para cada uno de ellos, el usuario debe especificar su duración, la franja del día a la que pertenecen (mañana, mediodía, tarde o noche) y unas ventanas de tiempo disponible (dentro de la cual el servicio debe realizarse) y óptima (dentro de la cual el usuario prefiere ser atendido). Las auxiliares tienen un contrato en el cual se especifica el número de horas que deben trabajar durante la semana, considerando como tiempo trabajado: el tiempo empleado realizando servicios, el tiempo empleado desplazándose entre las viviendas de los usuarios y todos los descansos que tengan durante el día (con la excepción del mayor de ellos si dura dos o más horas). Por otro lado, también se tienen unas afinidades que determinan cómo de adecuada es una auxiliar para realizar los servicios de un usuario.

El problema que queremos resolver es el que surge en la empresa cuando comienzan a trabajar en nueva ciudad o municipio. Ya que la única información disponible es la que proporcionan los usuarios, a los que se debe atender, y el objetivo de la empresa es determinar cuántas auxiliares deben contratar y sus planificaciones semanales.

Las planificaciones deben respetar las restricciones de las auxiliares (jornada diaria y semanal máxima permitida, existencia de sólo un descanso no remunerado, imposibilidad de realizar dos servicios a la vez) y usuarios (afinidades, ventanas de tiempo disponibles), así como las restricciones del problema de rutas (tiempos de inicio de servicio que respeten el desplazamiento entre usuarios), a la vez que se optimizan los siguientes objetivos:

1. Número de auxiliares que se utilizan.

2. Jornada trabaxada de las auxiliares.
3. Afinidad entre usuarios y auxiliares.
4. Ventanas optimas.

2. MÉTODO DE RESOLUCIÓN

Para resolver este problema de planificación lo hemos modelado como un problema de programación lineal entera, pero éste método solamente puede utilizarse para resolver problemas de pequeño tamaño. De modo que, para resolver problemas de tamaño real, se ha tenido que diseñar un algoritmo heurístico que se describe a continuación.

Planificación inicial

La planificación inicial (asignación de servicios a auxiliares en determinados horarios) se realiza de la forma siguiente:

1. Ordenar los servicios según su duración y ventana disponible.
2. Ordenar las auxiliares por afinidad.
3. Cada servicio se asigna a una auxiliar de modo que:
 - Se coloca en el horario más temprano posible.
 - Si la auxiliar no puede realizar el servicio, éste se asigna a la siguiente auxiliar.
4. A la solución se le aplica un postprocesado (i.e. optimización mediante una búsqueda local).

Optimización

La planificación inicial se optimiza (para reducir el número de auxiliares y la jornada trabaxada, así como para aumentar las afinidades y el número de servicios que se realizan dentro de sus ventanas óptimas) utilizando el método del *simulated annealing* de la siguiente forma:

1. Crear un vecindario a partir de la solución actual (aplicando movimientos que cambian el orden de los servicios o la auxiliar que los realiza) donde a cada vecino se le aplica un postprocesado (búsqueda local con movimientos que cambian el horario en que se realizan los servicios) con el objetivo de tratar de mejorar la jornada de las auxiliares.
2. El vecindario se utiliza para actualizar la solución actual siguiendo las normas clásicas del *simulated annealing*.

Movimientos

Los movimientos considerados durante la implementación del *simulated annealing*, i.e. movimientos que modifican la auxiliar que realiza los servicios o el orden en que son llevados a cabo, son:

Cambiar auxiliar Movimiento que consiste en cambiar la auxiliar que realiza un servicio.

Cambiar servicios de auxiliar Movimiento que consiste en cambiar de auxiliar un conjunto de varios servicios.

Intercambiar auxiliares Movimiento que consiste en intercambiar la auxiliar entre dos (o más) servicios.

Intercambiar orden Movimiento que consiste en intercambiar el orden en que se realizan varios servicios (asignados a la misma auxiliar en el mismo día).

Limpiar auxiliar Movimiento que consiste en eliminar todos los servicios que realiza una auxiliar (asignándolos a otra ptdiferente).

Los movimientos considerados durante el postprocesado, i.e. movimientos que modifican el horario en que se realizan los servicios, son:

Replanificar servicio Movimiento que consiste en cambiar el horario en que se realiza un servicio.

Desplazar servicios Movimiento que consiste en adelantar o atrasar el horario en que se realiza un conjunto de servicios consecutivos (asignados a la misma auxiliar en el mismo día).

Intercambiar horario Movimiento que consiste en intercambiar el horario de inicio entre dos servicios (asignados a la misma auxiliar en el mismo día).

Separar servicios Movimiento que consiste en crear un descanso entre dos servicios consecutivos.

Para comprobar el buen funcionamiento de este algoritmo realizamos un estudio computacional en el que comparamos los resultados obtenidos con los que proporciona Gurobi (para el problema de programación lineal entera). Analizando los resultados obtenidos vemos que nuestro algoritmo proporciona soluciones óptimas y, en algunos de los casos en que Gurobi no consigue resolver los ejemplos en un tiempo máximo de 1 día, se tiene que las soluciones que proporciona el algoritmo son mejores.

3. CONCLUSIONES

De modo que en este trabajo hemos estudiado un problema de planificación para una empresa de atención a domicilio y hemos presentado un algoritmo heurístico, basado en técnicas de búsqueda local, para su resolución.

REFERENCIAS

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671-680.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

OPTIMIZACIÓN WEB DUN SISTEMA DE XESTIÓN DE STOCKS

Luis Caviqúe¹, José Saias², Jorge Santos³, Armando B. Mendes⁴

¹ Universidade Aberta, BioISI-MAS luis.caviqúe@uab.pt

² Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Informática
jsaias@uevora.pt

³ Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática, Instituto
de Investigación e Formación Avanzada Centro de Investigación en Matemática e Aplicaciones,
jmas@uevora.pt

⁴ Universidade dos Açores, NIDeS e ALGORITMI armando.b.mendes@uac.pt

RESUMO

O obxectivo deste traballo é facilitar e acelerar o cálculo do nivel de orde e a súa cantidade óptima, nun sistema de xestión de stock con demanda e tempo de entrega aleatorios. Este cálculo era tradicionalmente laborioso e lento, xa que requiría un proceso iterativo de consulta de dúas táboas en cada iteración.

Polo tanto, o problema solucionouse inicialmente en follas de cálculo e posteriormente ampliouse a una aplicación en la Internet, onde o usuario fornece os datos e recibe os resultados.

Palabras e frases chave: Xestión de stock, Optimización, Probabilidade de escaseza, Fórmula de Wilson, Programación matemática.

1. INTRODUCCIÓN

Moitos grupos da industria utilizan a optimización do inventario para facer fronte á maior volatilidade e variabilidade da oferta, o que reduce o risco de sobrecarga e interrupción do inventario e optimiza o proceso de fabricación.

A optimización global dun sistema de stock de escaseza permitido con tempo de reposición e demanda aleatoria baixo o suposto de baixa probabilidade de escaseza leva a un proceso iterativo de consulta de táboas de función estándar normal e función normal de perda estándar como se describe na bibliografía (Liang et al, 2018).

A inicialización do proceso iterativo faise determinando a cantidade óptima sen rotura, tal e como se describe na fórmula de Economic Order Quantity, tamén coñecida como fórmula de Wilson (Tsan, 2013).

Dada a duración, lonxitude e dificultade deste proceso de cálculo manual, primeiro resolvemos o problema de minimización de custos empregando unha folla de cálculo e finalmente pasamos a resolver este problema na web.

O problema móstrase na figura 1 (tomada de Tavares et al, 2000) onde está representada a evolución do nivel de existencias na técnica de nivel de orde.

Esta técnica consiste en ordenar sempre a mesma cantidade Q sempre que o nivel de stock cae por debaixo dun limiar M chamado punto de pedido.

O obxectivo é determinar os valores de Q e M que minimizan o custo operativo do sistema, que consiste na suma de 3 porcións: custos da orde (minimizado ordenando grandes cantidades, pero raramente); Os custos de propiedade (minimizados ordenando pequenas cantidades a miúdo e con baixos niveis de orde) e, finalmente, os custos de reparto (minimizados mediante a adopción de niveis de orde elevada, mantendo sempre stocks elevados).

Existen conflitos entre os distintos obxectivos, polo que hai que abordar un problema de minimización non lineal.

Esta ferramenta de apoio á decisión foi desenvolvida con dous esquemas, sitio web e Java Swing (versión portátil que non require acceso a Internet). Empregando o método de desenvolvemento de

software coñecido como prototipado ou simplemente prototipo, foi posible desenvolver desde unha versión inicial da ferramenta ata o refinamento ata que se converteu na versión final. O uso deste método permitiu e facilitou a detección e mellora de fallos e funcionalidade das ferramentas.

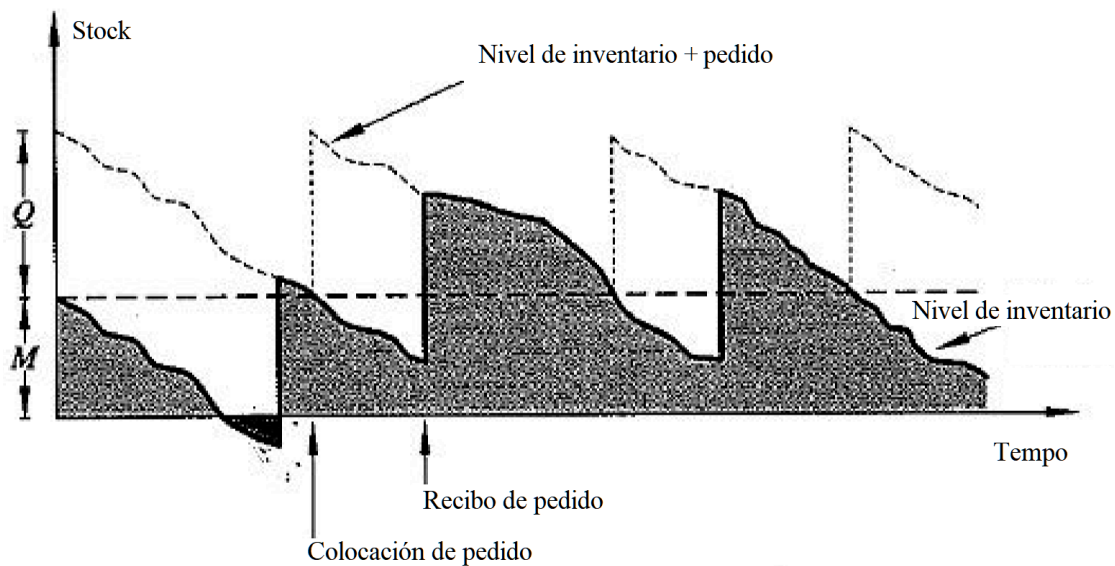


Figura 1: Evolución temporal dos niveis de existencias e cargas (Tavares et al 2000)

2. MÉTODOS

Os métodos utilizados consistiron principalmente en utilizar funcións estatísticas amplamente dispoñibles en follas de cálculo e linguaxes de programación usuais.

2.1. Descomposición da función de perda normal

Na literatura a función de perda normal emprégase para calcular o número medio de unidades que falta en cada ciclo. O número de unidades que faltan está representado por: η o que depende do consumo x durante o tempo de reposición e do nivel de orde M .

Polo tanto, temos a seguinte definición:

$$\eta(x, M) = \begin{cases} 0 & \text{se } x \leq M \\ x - M & \text{se } x > M \end{cases}$$

O número medio de unidades que falta por ciclo depende do nivel de orde M e da distribución da demanda durante o tempo de reposición $h(x)$

Polo tanto, dáse por:

$$\bar{\eta}(M) = \int_M^{+\infty} (x - M) h(x) dx$$

Na literatura, a distribución da demanda durante o tempo de reposición $h(x)$ é suposta normal con media μ e desviación estándar σ .

Recordando a expresión da respectiva función de densidade de probabilidade:

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Substituíndo na expresión do número medio de unidades que faltan por ciclo e normalizándose, esta pódese calcular a partir da función de densidade de probabilidade e da función de probabilidade acumulada, que están dispoñibles na maioría das follas de cálculo e nos principais idiomas de programación.

2.2. Implementación computacional

Inicialmente usamos follas de cálculo para confirmar tanto a velocidade de converxencia como a estabilidade do algoritmo.

3. CONCLUSIÓN

Descubriuse que, xeralmente, despois de aproximadamente 5 ou 6 iteracións converxeron os valores, na implementación web, tivemos que prestar especial atención a este problema, implementando un criterio de parada sempre que as solucións non presenten diferenzas por enriba dun limiar escollido.

Non obstante, lémbrese que a fórmula empregada só debe aplicarse a unha baixa probabilidade de rotura, é dicir, por baixo do 10%.

Comprobouse que un problema con iteracións e consultas tediosas a dúas táboas (Función normal de perda e Táboa de distribución normal) podería resolverse con facilidade nunha folla de cálculo. Coa facilidade de acceso a Internet hoxe en día, decidimos implementar o algoritmo nun servidor para permitir o acceso gratuito a esta ferramenta.

4. AGRADECEMENTOS

Os autores agradecen as súas respectivas universidades e centros de investigación pola súa colaboración.

REFERENCIAS

- Liang, Z.; Chaovalitwongse, W. A.; Shi, L. (2018) Supply Chain Management and Logistics: Innovative Strategies and Practical Solutions. CRC Press.
- Tsan-Ming, C. (2013) Handbook of EOQ Inventory Problems: Stochastic and Deterministic Models and Applications. Springer Science & Business Media
- Tavares, L.; Oliveira, R. C.; Correia, F. N. (2000). Investigação Operacional, McGraw-Hill,

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

PROBLEMA DE ASIGNACIÓN DE MUESTRAS DE LABORATORIO

Diego Noceda¹, Luisa Carpenté¹ y Silvia Lorenzo¹

¹ Grupo MODES. Universidade da Coruña

RESUMEN

El trabajo pretende resolver el problema de optimización asociado a la asignación de muestras de laboratorio en placas. El procesado de cada una de estas placas es costoso tanto en tiempo como en dinero, por tanto el principal objetivo es el de minimizar el número de placas utilizadas atendiendo a las restricciones impuestas por los termocicladores. Las características de este problema recuerdan al conocido Bin Packing, el cual es un problema de optimización NP-Hard que, si bien es factible modelar mediante un problema de programación lineal, no siempre va a ser posible encontrar la solución óptima con un coste razonable. Además, las restricciones añadidas sobre las muestras y la distribución de las placas llevan a un problema diferente. Así, el trabajo propone la implementación de un algoritmo que proporcione buenos resultados con un coste computacional bajo.

Palabras clave: Optimización, programación lineal entera, problemas de asignación, bin packing.

1. INTRODUCCIÓN

El trabajo diario de un laboratorio conlleva el procesamiento de placas llenas de muestras de ADN en termocicladores como parte de procesos más complejos. Debido al alto coste de procesar cada una de las placas tanto en tiempo como en dinero, el objetivo principal del trabajo es reorganizar las muestras de la mejor forma posible teniendo en cuenta las restricciones impuestas por las características del termociclador.

Las placas utilizadas por el laboratorio constan de 8 filas y 12 columnas, lo que hace un total de 96 pocillos donde colocar las muestras. Cada uno de estos pocillos contiene una mezcla entre una muestra de ADN y un componente químico que reacciona a cierta temperatura. Sin embargo, la selección de temperaturas para cada pocillo está limitada, ya que las placas se dividen en 6 franjas de temperatura, con 16 pocillos cada una, y la temperatura concreta se establece a nivel de franja completa. Además, por restricciones operativas del termociclador, la diferencia de temperatura entre dos franjas adyacentes debe ser inferior a 5° C.

Los datos de entrada del problema tienen las siguientes características, cada muestra ocupa un pocillo, un conjunto de muestras mezcladas con el mismo componente químico forman un grupo. Además, para garantizar el buen funcionamiento de los experimentos, para cada grupo presente en una placa es necesario reservar un pocillo para acomodar el control del grupo. La Figura 1 muestra una placa de ejemplo.

El objetivo del trabajo es obtener la distribución de las muestras en placas de forma que se minimice el número total de placas utilizadas, se minimice el número total de pocillos ocupados y se optimice el porcentaje de ocupación de las placas necesarias. Para la función objetivo se ha definido un orden lexicográfico para valorar las distintas soluciones.

2. ESTADO DEL ARTE

El problema, planteado por Carpenté et al (2019) ha sido resuelto utilizando un modelo de programación lineal entera, no factible en la práctica debido a su alto coste computacional, así como un algoritmo metaheurístico basado en el simulated annealing propuesto por Kirkpatrick

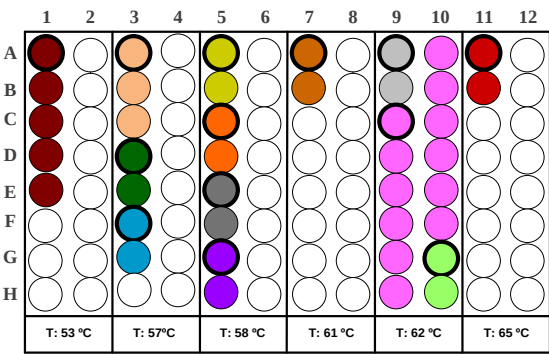


Figura 1: Ejemplo de placa con muestras colocadas en los pocillos.

et al (1983). Los resultados del algoritmo heurístico son buenos tanto en calidad como en coste computacional, pero presentan problemas de escalabilidad. Por tanto, es necesario dar un nuevo enfoque a la resolución del problema.

3. APROXIMACIÓN DIVIDE Y VENCERÁS

Se ha estudiado la estructura interna del problema y se ha diseñado una aproximación en 2 fases, en la primera el problema se realiza la descomposición del problema en otros más simples (aunque de complejidad variable). En una segunda fase las soluciones de todos estos subproblemas se combinan para crear una solución del problema completo. La ventaja de esta aproximación es que se puede mantener la optimalidad en cada una de las fases, gracias al desarrollo de dos modelos de programación lineal entera, lo que permite mejorar en gran medida la calidad de las soluciones, así como rebajar el coste computacional necesario.

REFERENCIAS

L. Carpen­te, A. Cerdeira-Pena, S. Lorenzo-Freire, A. S. Places. (2019) Optimization in Sanger sequencing. Computers & Operations Research, 109, 250–262.

S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing Science 1983, 220, 671-680.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

COOPERATIVE APPROACH TO A PLANT LOCATION PROBLEM WITH AGGLOMERATIVE ECONOMIES

Adriana Navarro Ramos¹ y Gustavo Bergantiños¹

¹ ECOSOT Economía, Sociedad y Territorio. Universidade de Vigo.

ABSTRACT

We study a plant location problem with agglomeration economies. In this problem, a firm that is planning to open a new plant in a country divided into regions. On these regions are located a finite number of firms that have some benefits from this incorporation. We assume that a planner should decide where to locate the new plant maximizing the global benefit. Then, a compensating scheme should be provided such that the incoming firm gets more than in the decentralized mechanism and the other firms are not worst off. In order to do this, we propose two rules to allocate the global benefit. Then, we approach the problem as a transferable-utility cooperative game. The core, the nucleolus and the τ -value of this game are studied and we analyze how the proposed rules relate to these concepts. Finally, we provide axiomatic characterizations for these rules.

Keywords: plant location, tu-game, core, nucleolus, τ -value.

1. INTRODUCTION

Firms dedicated to similar activities tend to settle their plants in the same region as others because to the proximity between them represents a mutual benefit. These benefits are called *agglomeration economies* and it is a well-studied phenomenon in the literature. The first analysis on this theory was made by Marshall (1920).

Consider that a firm is planning to open a new plant in certain country divided into different regions where the plant could be installed. There are also a finite number of firms already located in these regions and the new incorporation generates agglomeration economies to all of them.

In a decentralize mechanism, the plant would be located in the most profitable region for the newcomer firm. We assume that a welfare-maximizing social planner should decide the location of this new plant considering the total benefit of the country. Then, a compensation scheme should be propose in order to encourage the firm to locate the plant in the desired region.

Other important assumption is that the transfer of benefits necessary for the realization of such compensation scheme will be made exclusively through common resources (technologies, infrastructures, information, lands, etc.) shared among the firms within the same region. This means that all firms must be willing to cooperate so that the location and redistribution mechanisms can be carried out.

The first objective of this work is to model this situation and propose some rules to allocate the obtained global benefit. The second objective is to approach the problem as a transferable-utility cooperative game. We study three solution concepts: the core (Gillies, 1953), the nucleolus (Schmeidler, 1969) and the τ -value (Tijs, 1981). We also analyze how these solutions relate to the proposed rules. Finally, we provide axiomatic characterizations of the rules.

2. THE MODEL

An *agglomeration problem* is a tuple $\mathcal{A} = (0, N, P, b)$ where

- 0 is the firm which will open a plant in the country.

- $N = \{1, \dots, n\}$ is the set of firms already located in the country. We denote by $N_0 = N \cup \{0\}$.
- $R = \{1, \dots, r\}$ is the set of regions in the country and $P = \{P_1, \dots, P_r\}$ is a partition of N . P_k denotes the set of firms located in region k .
- $b = \{b_i^k : i \in N_0 = N \cup \{0\} \text{ and } k \in R\}$. $b_i^k \geq 0$ denotes the benefit obtained by firm i when 0 locates the plant at region k .

We assume that if an existing firm is outside the region where the new plant locates, this firm does not obtain any significant benefit, this is, for all $k \in R$ and all $i \notin P_k$, $b_i^k = 0$.

In a decentralized mechanism, firm 0 will be located in $\ell = \arg \max_{k \in R} \{b_0^k\}$. Instead, we determinate the optimal region and the global benefit of the problem.

Given a problem $\mathcal{A}$ we say that k^* is an *optimal location* if firm 0 should be located at region k^* in order to maximize the total profit. Namely for each $P_k \in P$,

$$\sum_{i \in P_{k^*} \cup \{0\}} b_i^{k^*} \geq \sum_{i \in P_k \cup \{0\}} b_i^k.$$

The *global benefit* of $\mathcal{A}$ is calculated as

$$g(\mathcal{A}) = \max_{k \in R} \left\{ \sum_{i \in N_0} b_i^k \right\}.$$

Finally, we define $L_0^{\mathcal{A}}$ as the maximum utility that firm 0 can obtain when cooperation is not achieved at region k^* . This is,

$$L_0^{\mathcal{A}} = \max_{k \in R} \left\{ \sum_{i \in N_0 \setminus P_{k^*}} b_i^k \right\}.$$

Note that $L_0^{\mathcal{A}}$ is the maximum between the individual benefit of firm 0 when it locates at k^* and the global benefit if firm 0 locates at the second most profitable region.

2. THE RULES

A *rule* is a function f assigning to each problem $\mathcal{A}$ a vector in $\mathbb{R}^{N_0}$ satisfying that

$$\sum_{i \in N_0} f_i(\mathcal{A}) = g(\mathcal{A}).$$

For every problem $\mathcal{A}$ and each $i \in N_0$, the *egalitarian optimal location rule* is given by

$$EOL_i(\mathcal{A}) = \begin{cases} L_0^{\mathcal{A}} + \frac{g(\mathcal{A}) - L_0^{\mathcal{A}}}{|P_{k^*}| + 1}, & \text{if } i = 0 \\ \frac{g(\mathcal{A}) - L_0^{\mathcal{A}}}{|P_{k^*}| + 1}, & \text{if } i \in P_{k^*} \\ 0, & \text{otherwise.} \end{cases}$$

For each problem $\mathcal{A}$ and each $i \in N_0$, the *weighted optimal location rule* is

$$WOL_i(\mathcal{A}) = \begin{cases} L_0^{\mathcal{A}} + \frac{b_0^{k^*}}{g(\mathcal{A})}(g(\mathcal{A}) - L_0^{\mathcal{A}}), & \text{if } i = 0 \\ \frac{b_i^{k^*}}{g(\mathcal{A})}(g(\mathcal{A}) - L_0^{\mathcal{A}}), & \text{if } i \in P_{k^*} \\ 0, & \text{otherwise.} \end{cases}$$

These rules have a nice interpretation. Firm 0 receives $L_0^{\mathcal{A}}$ for sure. Then, the surplus generated (with respect to $L_0^{\mathcal{A}}$) by P_{k^*} and 0 is divided. In *EOL* equally among all firms generating such surplus.

The difference between these rule is that, the surplus is divided equally among all agents in $P_{k^*} \cup \{0\}$ whereas in *WOL*, $g(\mathcal{A}) - L_0^{\mathcal{A}}$ is divided in proportion to b .

3. THE TU-GAME

We now define $v^{\mathcal{A}}$, a characteristic function reflecting the assumption that transference of benefits with firm 0 can only take place if all firms already installed at a region k are willing to cooperate.

For all $S \subseteq N_0$, we define the set

$$U(S) = \{i \in S : P_{k(i)} \subseteq S\},$$

where $k(i) \in R$ is the region where the firm i is located. This set indicates the firms whose region is contained in S .

Then, for any problem $\mathcal{A}$ we define the characteristic function

$$v^{\mathcal{A}}(S) = \begin{cases} \max_{k \in R} \left\{ b_0^k + \sum_{i \in U(S)} b_i^k \right\}, & \text{if } 0 \in S \\ 0, & \text{otherwise,} \end{cases}$$

for every $S \subseteq N_0$. Since $v^{\mathcal{A}}(N_0) = g(\mathcal{A})$, this characteristic function can be used to model an agglomeration problem.

4. SOLUTIONS IN TU-GAMES

We first study the *core* of $v^{\mathcal{A}}$. The core of a game consist of all stable allocations; this means no coalition can improve upon a core element by splitting off.

We prove that the core of $(N_0, v^{\mathcal{A}})$ is always non empty and it is given by

$$C(N_0, v^{\mathcal{A}}) = \left\{ x \in \mathbb{R}^{n+1} : \sum_{i \in N_0} x_i = g(\mathcal{A}), L_0^{\mathcal{A}} \leq x_0 \leq g(\mathcal{A}), \right. \\ \left. 0 \leq x_i \leq g(\mathcal{A}) - L_0^{\mathcal{A}}, \forall i \in P_{k^*}, x_i = 0, \forall i \notin P_{k^*} \right\}.$$

Next, we study two single-value solutions: *nucleolus* and τ -value.

The nucleolus is a lexicographical minimum over all vectors that divide the joint profits in such a way that every player receives at least the profit she can achieve on her own. The τ -value is a compromise between an upper bound and a lower bound on what the players can demand from the joint profits.

We prove that for any problem $\mathcal{A}$, the nucleolus and the τ -value of game (N_0, v) coincide with the egalitarian optimal location rule.

5. CHARACTERIZATIONS

In this section we introduce some properties of a rule for any problem $\mathcal{A}$ and the axiomatic characterizations for the optimal location rule and the weighted optimal location rule.

Core selection. A rule gives an allocation belonging to the core of the game $(N_0, \mathcal{A})$.

Equal treatment. If the value of the optimal location increases, then all firms of such region and firm 0 should gain the same.

Weighted treatment. If the value of the optimal location increases, then all firms of such region and firm 0 should gain in proportion of the value they generate on the original problem.

The egalitarian optimal location rule is the unique rule satisfying core selection and equal treatment.

The weighted optimal location rule is the unique rule satisfying core selection and weighted treatment.

REFERENCES

- Gillies, D. (1953) Some theorems on n-person games. Ph. D. Dissertation, Princeton University, Department of Mathematics.
- Marshall, A. (1920) Principles of Economics. London: MacMillan, 8 edition.
- Schmeidler D. (1969) The nucleolus of a characteristic function game. SIAM Journal on applied mathematics, 17(6), 1163-1170.
- Tijs S. (1981) Bounds for the core of a game and the t-value. Game Theory and Mathematical Economics, 123-132.

UNCERTAINTY IN COOPERATIVE INTERVAL GAMES: HOW HURWICZ CRITERION COMPATIBILITY LEADS TO EGALITARIANISM

Lina Mallozzi¹ and Juan Vidal-Puga²

¹Dipartimento di Matematica e Applicazioni, Università di Napoli Federico II, Via Claudio 21, 80125 Naples, Italy.

²Department of Statistics and Operations Research, Universidade de Vigo, 36200 Vigo (Pontevedra), Spain.

ABSTRACT

We study cooperative interval games. These are cooperative games where the value of a coalition is given by a closed real interval specifying a lower bound and an upper bound of the possible outcome. For interval cooperative games, several (interval) solution concepts have been introduced in the literature. We assume that each player has a different attitude towards uncertainty by means of the so-called Hurwicz coefficients. These coefficients specify the degree of optimism that each player has, so that an interval becomes a specific payoff. We show that a classical cooperative game arises when applying the Hurwicz criterion to each interval game. On the other hand, the same Hurwicz criterion can be also applied to any interval solution of the interval cooperative game. Given this, we say that a solution concept is Hurwicz compatible if the two procedures provide the same final payoff allocation. When such compatibility is possible, we characterize the class of compatible solutions, which reduces to the egalitarian solution when symmetry is required. The Shapley value and the core solution cases are also discussed.

Keywords: Game theory, cooperative interval games, Hurwicz criterion, Hurwicz compatibility.

1. INTRODUCTION

Given a set of agents (or players), cooperative (or transferable utility) games assign to each coalition of agents a real number which represents the maximum utility that the members of this coalition can assure by themselves. Cooperative game theory has addressed these problems by proposing relevant solutions, or values, that suggest one or several payoff allocations satisfying certain desirable properties. Typically, the most standard property is efficiency, which implies that the worth of the grand coalition is shared. Some examples of efficient solutions are the core, the Weber set, the Shapley value, and the nucleolus.

Cooperative interval games generalize the idea of cooperative games by assigning to each coalition a closed interval. Analogously, efficient interval solutions propose an allocation of the interval generated by the grand coalition. Many of the classical solutions have been defined in the context of cooperative interval games by [2, 5]. See [14] for a survey.

Interval games have been applied to bankruptcy problems [13], airport games [3], minimum cost spanning tree problems [25, 26], assignment problems [27, 32, 33], and sequencing games [6].

A classical interpretation of the intervals is that each of them represents the possible worth range that a coalition can get by themselves. Examples are those that appear from the so-called games with externalities [29], where the worth of a coalition depends not only on the coalition itself, but also on how the rest of the players cooperate.¹ [32] propose a different motivation, where the worth of a coalition varies between the (classical) pessimistic assumption that the rest of players will try to harm them as much as possible, and the most optimistic assumption given by the dual problem.

¹For example, assume that players form an oligopoly that plans to create a cartel. The cartel can then anticipate their benefit as a monopoly. However, if two or more players are not present, the remaining players can not anticipate their exact benefit, as it would depend on whether the other players merge or not.

At this point, we have to distinguish between risk and uncertainty. In a risky situation, players are unsure of the final result of their own cooperation, but they can assign a precise probability to each possible outcome. This kind of riskiness has been deeply studied in the economic literature, both from a cooperative and non-cooperative point of view. Frequently, it is assumed that each player has some private information and a precise knowledge of the probability distribution of how the others are. Yet, this approach does not fit into the model proposed by interval games, where each interval does not depend on the private information of the players. In other situations, probability distribution is common knowledge but their consequences are not homogeneous among players [6]. Again, this approach neither fits interval games, where each interval is coalition-dependent.

As opposed, under uncertainty, players are not only unsure of the final outcome of their potential cooperation, but also of the probability of these possible outcomes. When there is no private information, i.e. all the players agree on the uncertainty that lies behind the cooperation of each coalition, interval games provide a more realistic interpretation of uncertainty.

An interval solution is then a way to share the uncertainty of the grand coalition worth taking into account the uncertain worth of each coalition.

Given this, each player can have a different attitude towards uncertainty. For example, a pessimistic player would prefer to maximize the minimum possible outcome (*maximin criterion*), so that its preferred payoff is an interval with a high lower bound, whereas an optimistic (*maximax criterion*) player would prefer intervals with a high upper bound. An intermediate approach (*Laplace criterion* or *criterion of rationality*) is to assume all the possible outcomes are equally probable, in the sense that they follow an uniform distribution. Hence, players with a Laplace criterion prefer intervals with a high midpoint.

A generalization of these criteria is the so-called Hurwicz criterion [20], which states that there exists a fixed coefficient between 0 and 1 that measures the degree of optimism. Hence, a pessimistic player would have a coefficient 0, an optimistic one 1, and a rational one (in the sense of Laplace) $\frac{1}{2}$.

Obviously, once the Hurwicz coefficients are stated for each player, uncertainty disappears and players can uniquely assign a concrete value to each interval. The payoff allocation of the grand coalition interval becomes a payoff allocation for the grand coalition. We can then check the payoff allocation proposed by each interval solution.

In this paper, we study what happens when this statement of the coefficients is done on the interval game, prior to applying any interval value. We prove (Proposition 1) that this operation generates a (classical) cooperative game, so that we can compute its interval value.²

Hence, we can proceed in two ways. On the one hand, we can compute the interval value on the interval game, and then apply the Hurwicz criterion. On the other hand, we can apply the Hurwicz criterion in the interval game in order to get a (classical) cooperative game, then apply the interval solution.³ We are interested in studying which values are compatible in the sense that both procedures provide the same final payoff allocation. This property has the potential of solving situations where players have uncertain needs for a resource when it has to be divided before uncertainty resolves [34].

Our second result (Proposition 2) implies that this property is incompatible with efficiency in the more general setting. However, we focus on the only two situations where they are compatible. On one hand, we can assume that all the players have the same Hurwicz coefficient. On the other hand, we can assume that uncertainty disappears when the grand coalition forms. The latter is what actually happens in the motivating examples: both games with externalities and the optimistic-pessimistic approach given by [32].

We prove (Theorem 1 and Theorem 3) that the only compatible solutions are the proportional ones. These are solutions that have recently received increasing attention from the economic and social choice theory [1, 8, 22, 35]. They state that the worth of the grand coalition should be proportionally distributed. As a direct corollary, we deduce (Theorem 2 and Theorem 4) that the unique anonymous (in the sense that symmetric agents are treated symmetrically) compatible solution is the egalitarian one. This states an equivalence between Hurwicz compatibility and the egalitarian notion given by the Dutta-Ray [18] interpretation ideally yielding Lorenz-dominant allocations but without core-like participation constraints. Egalitarian outcomes are also backed by experimental evidence [12]. It is worthy to note that usual

²Other applications of Hurwicz coefficients in interval games appear in [23] and [24], who also deduce a (classical) cooperative game by using a selection via degrees of optimism. However, these degrees are coalition-dependent, not individual. Hence, they cannot be identified as Hurwicz coefficients in the same way we do here.

³This is always possible to do, since interval cooperative games generalize classical ones.

characterizations of the egalitarian solution heavily rely on the properties of either additivity [8, 9, 10, 17, 21, 30, 31] or monotonicity [11, 15, 16], which are not required here.

The rest of the paper is organized as follows. In Section 2, we present the notation. In Section 3, we define the Hurwicz criterion, we prove that it can only be applied in two particular situations, which we analyze, and we study the core. In Section 4, we present some concluding remarks.

2. NOTATION

Let $\mathbb{I} := \{ia = [\underline{ia}, \bar{ia}] : \underline{ia}, \bar{ia} \in \mathbb{R}, \underline{ia} \leq \bar{ia}\}$ be the set of closed intervals in $\mathbb{R}$. Given $ia, ib \in \mathbb{I}$, we say that $ia \preceq ib$ when $\underline{ia} \leq \underline{ib}$ and $\bar{ia} \leq \bar{ib}$. Notice that if $ia = [a, a]$ and $ib = [b, b]$ for some $a, b \in \mathbb{R}$, we have that $ia \preceq ib$ iff $a \leq b$. Remark that $\preceq$ is a partial order relation in $\mathbb{I}$; for example $[-1, 2] \preceq [0, 2]$ but $[-2, 3]$ and $[-1, 2]$ are not comparable with respect to $\preceq$.

Given $x \in \mathbb{R}_+$ and $iy = [\underline{iy}, \bar{iy}] \in \mathbb{I}$, we define $x \cdot iy := [x \cdot \underline{iy}, x \cdot \bar{iy}] \in \mathbb{I}$. Given $\alpha_j \in \mathbb{R}$ and $ia = [\underline{ia}, \bar{ia}] \in \mathbb{I}$, we define

$$\alpha_j \circ ia := \alpha_j \cdot \bar{ia} + (1 - \alpha_j) \cdot \underline{ia} = \underline{ia} + \alpha_j \cdot (\bar{ia} - \underline{ia}) \in \mathbb{R}.$$

Given $ia, ib \in \mathbb{I}$, we define $ia \oplus ib := [\underline{ia} + \underline{ib}, \bar{ia} + \bar{ib}] \in \mathbb{I}$, and $ia \odot ib := [\underline{ia} - \bar{ib}, \bar{ia} - \underline{ib}] \in \mathbb{I}$; moreover, when $\underline{ia} - \underline{ib} \leq \bar{ia} - \bar{ib}$, we define $ia \ominus ib := [\underline{ia} - \underline{ib}, \bar{ia} - \bar{ib}] \in \mathbb{I}$. When $iA = \{ia_1, \dots, ia_k\} \subset \mathbb{I}$, we define

$$\sum_{ia \in iA} ia := ia_1 \oplus \dots \oplus ia_k \in \mathbb{I}$$

with the convention that $\sum_{ia \in \emptyset} ia := [0, 0]$.

Let $N = \{1, \dots, n\}$ be a finite set of players. A *coalitional interval game* is a pair (N, iv) where $iv : 2^N \rightarrow \mathbb{I}$ is a function that assigns a closed interval $iv(S) = [\underline{iv}(S), \bar{iv}(S)]$ to each coalition $S \subseteq N$ with the property that $iv(\emptyset) = [0, 0]$, i.e. $\underline{iv}(\emptyset) = \bar{iv}(\emptyset) = 0$. Let $\mathcal{IG}^N$ denote the class of all coalitional interval games with N as set of players. Since N is fixed, from now on we write iv instead of (N, iv) and $\mathcal{IG}$ instead of $\mathcal{IG}^N$.

Notice that coalitional interval games generalize classical coalitional (*transferable utility*, or TU) games. Just take $\underline{iv}(S) = \bar{iv}(S)$ for all $S \subseteq N$. There are three trivial TU games associated to any $iv \in \mathcal{IG}$. These are the border games $\bar{iv}$ and $\underline{iv}$, and the length game $|iv|$ given by $|iv|(S) = \bar{iv}(S) - \underline{iv}(S)$ for all $S \subseteq N$. Let $\mathcal{G}$ denote the set of TU games with N as player set. With some abuse of notation, we assume $\mathcal{G} \subset \mathcal{IG}$.

Apart from coalitional interval games, another generalization of TU games are *non-transferable utility games*, or NTU games. An NTU game V with player set N is given by a characteristic function $V : 2^N \rightarrow \bigcup_{S \subseteq N} \mathbb{R}^S$, with the convention $\mathbb{R}^\emptyset = \{0\}$, satisfying, for all $S \subseteq N$:

1. $V(S) \subseteq \mathbb{R}^S$
2. $V(S) \neq \emptyset$
3. $V(S)$ closed
4. $V(S)$ bounded from above, i.e. $\exists x \in \mathbb{R}^S$ such that $V(S) \cap \{y \in \mathbb{R}^S : y \geq x\} = \emptyset$
5. $V(S)$ comprehensive, i.e. $x \in V(S), y \leq x \implies y \in V(S)$.

Any TU game v can be written as an NTU game V as follows:

$$V(S) = \left\{ x \in \mathbb{R}^S : \sum_{i \in S} x_i \leq v(S) \right\}$$

for all $S \subseteq N$.

Notice that we can write any interval game $iv \in \mathcal{IG}$ applying natural generalization of NTU games as follows:

$$iV(S) = \left\{ ia \in \mathbb{I}^S : \sum_{j \in S} ia_j \preceq iv(S) \right\} \quad (1)$$

for all $S \subseteq N$. We then say that iV is an *interval game written in NTU form*.

A relevant class of coalitional interval games is the following [4]. A coalitional interval game iv is *size monotonic* if $|iv|(S) \leq |iv|(T)$ for all $S \subseteq T \subseteq N$. We denote as $\mathcal{SMIG}$ the set of size monotonic interval game with N as player set. Clearly, all TU games are size monotonic interval games, i.e. $\mathcal{G} \subset \mathcal{SMIG}$.

A *solution in the set of TU games* is a function $\sigma : \mathcal{G} \rightarrow \mathbb{R}^N$ that assigns to each coalitional TU game $v \in \mathcal{G}$ a payoff allocation $\sigma(v) \in \mathbb{R}^N$. A solution σ in the set of coalitional games is *efficient* if $\sum_{j \in N} \sigma_j(v) = v(N)$ for all $v \in \mathcal{G}$. A well-known efficient solution in the set of TU games is the *Shapley value* [28].

Let $\mathcal{IB}$ be a subset of $\mathcal{IG}$ such that it contains all the TU games, i.e. $\mathcal{G} \subseteq \mathcal{IB}$. An *efficient solution for $\mathcal{IB}$* is a function $i\sigma : \mathcal{IB} \rightarrow \mathbb{I}^N$ that assigns to each $iv \in \mathcal{IB}$ a payoff allocation $i\sigma(iv) \in \mathbb{I}^N$ such that $\sum_{j \in N} i\sigma_j(iv) = iv(N)$ for all $iv \in \mathcal{IB}$.

3. HURWICZ CRITERION

The most well-known criterion to deal with uncertainty was first stated by Hurwicz in [20]. Assume that each player $j \in N$ has a coefficient $\alpha_j \in [0, 1]$ which determines its degree of optimism. This means that if player j faces an interval $ia \in \mathbb{I}$, its individual valuation of it will be $\alpha_j \circ ia$.

Given $\alpha \in [0, 1]^N$ and an interval game solution $i\sigma : \mathcal{IB} \rightarrow \mathbb{I}^N$, we define $\alpha \circ i\sigma : \mathcal{IB} \rightarrow \mathbb{R}^N$ as the function given by applying the Hurwicz criterion to $i\sigma$ with coefficients in α , i.e.

$$(\alpha \circ i\sigma)_j(iv) := \alpha_j \circ i\sigma_j(iv) \in \mathbb{R} \quad (2)$$

for all $iv \in \mathcal{IB}$ and all $j \in N$.

Analogously, given $\alpha \in [0, 1]^N$ and $iv \in \mathcal{IG}$, we define $\alpha \circ iV$ as the NTU game given by applying the Hurwicz criterion to each $iv(S)$ with coefficients in α . Formally,

$$(\alpha \circ iV)(S) := \left\{ (\alpha_j \circ ia_j)_{j \in S} \in \mathbb{R}^S : ia \in iV(S) \right\} - \mathbb{R}_+^S \quad (3)$$

for all $S \subseteq N$, where iV is defined as in (1). As usual, $-\mathbb{R}_+^S$ allows us to assure comprehensivity (despite the incompleteness of the partial order $\preceq$).

Proposition 1 *Given $\alpha \in [0, 1]^N$ and $iv \in \mathcal{IG}$, the associated NTU game $\alpha \circ iV$ is equivalent to the TU game $\alpha \circ iv$ defined as follows:*

$$(\alpha \circ iv)(S) := \left(\max_{j \in S} \alpha_j \right) \circ iv(S) \in \mathbb{R} \quad (4)$$

for all $S \subseteq N$, $S \neq \emptyset$.

Proof. Fix $S \subseteq N$, $S \neq \emptyset$. If $S = \{j\}$, then

$$\begin{aligned} (\alpha \circ iV)(S) &= (\alpha \circ iV)(\{j\}) \stackrel{(3)(1)}{=} \{ \alpha_j \circ ia_j : ia_j \preceq iv(\{j\}) \} - \mathbb{R}_+^{\{j\}} \\ &= \{ \alpha_j \circ iv(\{j\}) \} - \mathbb{R}_+^{\{j\}} = \left\{ x \in \mathbb{R}^{\{j\}} : x \leq \alpha_j \circ iv(\{j\}) \right\} \\ &\equiv \alpha_j \circ iv(\{j\}) = \left(\max_{j \in S} \alpha_j \right) \circ iv(S). \end{aligned}$$

Take $S \subseteq N$, $|S| > 1$, and $k \in \arg \max_{j \in S} \alpha_j$. In case of more than one possible k , we take any one of

them. Let $T = S \setminus \{k\} \neq \emptyset$. For each $ia \in \mathbb{I}$, let $|ia| := \overline{ia} - \underline{ia}$. Then,

$$\begin{aligned}
 & (\alpha \circ iV)(S) \\
 & \stackrel{(3)}{=} \left\{ (\alpha_j \circ ia_j)_{j \in S} \in \mathbb{R}^S : ia \in iV(S) \right\} - \mathbb{R}_+^S \\
 & \stackrel{(1)}{=} \left\{ (\alpha_j \circ ia_j)_{j \in S} \in \mathbb{R}^S : \sum_{j \in S} ia_j \preceq iv(S) \right\} - \mathbb{R}_+^S \\
 & = \left\{ (\alpha_j \circ ia_j)_{j \in S} \in \mathbb{R}^S : \sum_{j \in S} ia_j = iv(S) \right\} - \mathbb{R}_+^S \\
 & = \left\{ (x_j + \alpha_j y_j)_{j \in S} \in \mathbb{R}^S : \sum_{j \in S} x_j = \underline{iv}(S), x \in \mathbb{R}^S, \sum_{j \in S} y_j = |iv|(S), y \in \mathbb{R}_+^S \right\} - \mathbb{R}_+^S.
 \end{aligned}$$

Given that $x \in \mathbb{R}^S$ is only restricted by $\sum_{j \in S} x_j = \underline{iv}(S)$, this amount $\underline{iv}(S)$ can be freely transferable among the players, so that the Pareto frontier is reached when $\sum_{j \in S} \alpha_j y_j$ is maximum, given that $y \in \mathbb{R}_+^N$ and $\sum_{j \in S} y_j = |iv|(S)$. Since $\alpha_k \geq \alpha_j$ for all $j \in S$, this maximum is $\alpha_k \cdot |iv|(S)$, reached at least when $y_k = |iv|(S)$ and $y_j = 0$ otherwise. Hence,

$$\begin{aligned}
 (\alpha \circ iV)(S) & = \left\{ (x_j)_{j \in T} \times (x_k + \alpha_k \cdot |iv|(S)) \in \mathbb{R}^S : \sum_{j \in S} x_j = \underline{iv}(S) \right\} - \mathbb{R}_+^S \\
 & = \left\{ x \in \mathbb{R}^S : \sum_{j \in S} x_j = \underline{iv}(S) + \alpha_k |iv|(S) \right\} - \mathbb{R}_+^S \\
 & \equiv \alpha_k \circ iv(S) = \left(\max_{j \in S} \alpha_j \right) \circ iv(S).
 \end{aligned}$$

□ ■

Under Proposition 1, any coalitional interval game $iv \in \mathcal{IG}$ turns into a unique TU game $\alpha \circ iv \in \mathcal{G}$ by applying the Hurwicz criterion with coefficients in α .

Given $\alpha \in [0, 1]^N$ and an interval game solution $i\sigma : \mathcal{IB} \rightarrow \mathbb{I}^N$, we define $i\sigma \circ \alpha : \mathcal{IB} \rightarrow \mathbb{R}^N$ as the function given by applying $i\sigma$ to $\alpha \circ iv$ for each $iv \in \mathcal{IB}$, i.e.

$$(i\sigma \circ \alpha)_j(iv) := i\sigma_j(\alpha \circ iv) \in \mathbb{R} \quad (5)$$

for all $iv \in \mathcal{IB}$ and $j \in N$.

Notice that both (2) and (5) apply an interval solution and the Hurwicz criterion with some coefficients. The difference between both approaches is the order in which they do so.

It is then natural to require this order to be irrelevant, i.e. both $\alpha \circ i\sigma$ and $i\sigma \circ \alpha$ should coincide. We call this property *Hurwicz compatibility*. Formally,

Definition 1 An interval game solution $i\sigma : \mathcal{IB} \rightarrow \mathbb{I}^N$ is Hurwicz compatible in $\mathcal{IB}$ if $(i\sigma \circ \alpha)(iv) = (\alpha \circ i\sigma)(iv)$ for all $iv \in \mathcal{IB}$ and all $\alpha \in [0, 1]^N$.

Proposition 2 For all $\alpha \in [0, 1]^N$ and all $iv \in \mathcal{IG}$,

$$\sum_{j \in N} (\alpha \circ i\sigma)_j(iv) \leq \sum_{j \in N} (i\sigma \circ \alpha)_j(iv)$$

for all efficient solution $i\sigma : \mathcal{IG} \rightarrow \mathbb{I}^N$. Moreover, equality holds iff either

- $\alpha_j = \alpha_k$ for all $j, k \in N$ or
- $|iv|(N) = 0$.

Proof. Under efficiency,

$$\begin{aligned}
 \sum_{j \in N} (\alpha \circ i\sigma)_j(iv) &\stackrel{(2)}{=} \sum_{j \in N} \alpha_j \circ i\sigma_j(iv) = \sum_{j \in N} (\underline{i\sigma_j(iv)} + \alpha_j \cdot |i\sigma_j(iv)|) \\
 &= \underline{iv}(N) + \sum_{j \in N} \alpha_j |i\sigma_j(iv)| \leq \underline{iv}(N) + \left(\max_{j \in N} \alpha_j \right) \sum_{j \in N} |i\sigma_j(iv)| \\
 &= \underline{iv}(N) + \left(\max_{j \in N} \alpha_j \right) |iv|(N) = \left(\max_{j \in N} \alpha_j \right) \circ iv(N) \\
 &\stackrel{(4)}{=} (\alpha \circ iv)(N) = \sum_{j \in N} i\sigma_j(\alpha \circ iv) \stackrel{(5)}{=} \sum_{j \in N} (i\sigma \circ \alpha)_j(iv).
 \end{aligned}$$

Equality holds iff

$$\sum_{j \in N} \alpha_j |i\sigma_j(iv)| = \left(\max_{j \in N} \alpha_j \right) \sum_{j \in N} |i\sigma_j(iv)|$$

which is equivalent to either $\alpha_j = \alpha_k$ for all $j, k \in N$ or $|i\sigma_j(iv)| = 0$ for all $j \in N$. Since $|x|$ is always nonnegative, this second condition is equivalent to $|iv|(N) = 0$. $\square$ ■

Corollary 1 *No efficient solution in the set of (size monotonic) coalitional interval games is Hurwicz compatible.*

In view of Proposition 2, we can only find Hurwicz compatibility in efficient solutions when we restrict ourselves to two possible situations:

1. All the Hurwicz coefficients coincide (equal degree of optimism).
2. There is no uncertainty when all players cooperate (grand coalition certainty).

We study which interval solutions are Hurwicz compatible when all coefficients coincide.

Definition 2 *Given $\mathcal{IB} \subseteq \mathcal{IG}$, an interval game solution $i\sigma : \mathcal{IB} \rightarrow \mathbb{I}^N$ is uniform Hurwicz compatible in $\mathcal{IB}$ if, for all $\alpha_0 \in [0, 1]$, we have $(i\sigma \circ \alpha)(iv) = (\alpha \circ i\sigma)(iv)$ for all $iv \in \mathcal{IB}$, where $\alpha_j = \alpha_0$ for all $j \in N$.*

Let us consider a coalitional interval game iv and $\alpha \in [0, 1]^N$ the vector of players' Hurwicz coefficients. We denote $\alpha_S := \max_{j \in S} \alpha_j$ and

$$n_S := \frac{|S|!(n-1-|S|)!}{n!}$$

for each $S \subseteq N$. By convention, $\alpha_\emptyset := 0$.

We consider here the Shapley value of the TU game $\alpha \circ iv$, i.e.

$$Sh_j(\alpha \circ iv) := \sum_{S, j \notin S} n_S (\alpha_{S \cup \{j\}} \circ iv(S \cup \{j\}) - \alpha_S \circ iv(S)) \quad (6)$$

for each $j \in N$. On the class of size monotonic coalitional interval games, Alparslan Gök, Brânzei and Tijs [7] define a generalization of the Shapley value, which we call the *interval Alparslan Gök-Brânzei-Tijs (ABT) solution*, as follows:

$$iABT_j(iv) := \sum_{S, i \notin S} n_S \cdot (iv(S \cup \{i\}) \ominus iv(S))$$

for each $iv \in \mathcal{SMIG}$ and each $j \in N$. Analogously, Han, Sun and Xu [19] define another generalization of the Shapley value, which we call the *interval Han-Sun-Xu (HSX) solution*, as follows:

$$iHSX_j(iv) := \sum_{S, i \notin S} n_S \cdot (iv(S \cup \{i\}) \oslash iv(S))$$

for each $iv \in \mathcal{IG}$ and each $j \in N$.

Proposition 3 *The interval ABT solution is uniform Hurwicz compatible in $SMIG$.*

Proof. The interval ABT solution can be written as

$$iABT_j(iv) = \left[\sum_{S, j \notin S} n_S (\underline{iv}(S \cup \{j\}) - \underline{iv}(S)), \sum_{S, j \notin S} n_S (\overline{iv}(S \cup \{j\}) - \overline{iv}(S)) \right]$$

for each $iv \in SMIG$ and each $j \in N$. Since all the Hurwicz coefficients coincide ($\alpha_1 = \alpha_2 = \dots = \alpha_n = \alpha_0$), then $\alpha \circ iABT(iv) = Sh(\alpha \circ iv)$. In particular,

$$\begin{aligned} & \alpha \circ iABT_j(iv) \\ &= (1 - \alpha_0) \left(\sum_{S, j \notin S} n_S (\underline{iv}(S \cup \{j\}) - \underline{iv}(S)) \right) \\ &+ \alpha_0 \left(\sum_{S, j \notin S} n_S (\overline{iv}(S \cup \{j\}) - \overline{iv}(S)) \right) \\ &= \sum_{S, j \notin S} n_S ((1 - \alpha_0) (\underline{iv}(S \cup \{j\}) - \underline{iv}(S)) + \alpha_0 (\overline{iv}(S \cup \{j\}) - \overline{iv}(S))) \\ &= \sum_{S, j \notin S} n_S (((1 - \alpha_0) \underline{iv}(S \cup \{j\}) + \alpha_0 \overline{iv}(S \cup \{j\})) - ((1 - \alpha_0) \underline{iv}(S) + \alpha_0 \overline{iv}(S))) \\ &= \sum_{S, j \notin S} n_S (\alpha_0 \circ iv(S \cup \{j\}) - \alpha_0 \circ iv(S)) = Sh_j(\alpha \circ iv) \\ &= iABT_j(\alpha \circ iv) = (iABT \circ \alpha)_j(iv) \end{aligned}$$

for each $iv \in SMIG$ and each $j \in N$. □ ■

Theorem 1 *An efficient solution $i\sigma : \mathcal{IG} \rightarrow \mathbb{R}^N$ is uniform Hurwicz compatible in $\mathcal{IG}$ if and only if there exist two vectors $\delta \in \mathbb{R}^N$ and $\gamma \in \mathbb{R}_+^N$ with $\sum_{j \in N} \delta_j = 0$ and $\sum_{j \in N} \gamma_j = 1$, and such that*

$$i\sigma_j(iv) = \delta_j \cdot [1, 1] + \gamma_j \cdot iv(N) \quad (7)$$

for all $iv \in \mathcal{IG}$ and all $j \in N$.

Proof. ($\Leftarrow$) Assume $i\sigma$ is defined as in (7). It is clear that such a solution is efficient. We now check that it is uniform Hurwicz compatible. Fix $iv \in \mathcal{IG}$. Let $\alpha_0 \in [0, 1]$ and let $\alpha \in [0, 1]^N$ defined as $\alpha_j = \alpha_0$ for all $j \in N$. We have to prove that $(i\sigma \circ \alpha)(iv) = (\alpha \circ i\sigma)(iv)$. For each $j \in N$,

$$\begin{aligned} (i\sigma \circ \alpha)_j(iv) &= i\sigma_j(\alpha \circ iv) \\ &= \delta_j \cdot [1, 1] + \gamma_j \cdot (\alpha \circ iv(N)) \\ &\equiv \delta_j + \gamma_j \cdot (\alpha \circ iv(N)) \end{aligned}$$

and

$$\begin{aligned} (\alpha \circ i\sigma)_j(iv) &= \alpha \circ i\sigma_j(iv) \\ &= \alpha \circ (\delta_j \cdot [1, 1] + \gamma_j \cdot iv(N)) \\ &= \delta_j + \gamma_j \cdot (\alpha \circ iv(N)). \end{aligned}$$

($\Rightarrow$) Let $i\sigma : \mathcal{IG} \rightarrow \mathbb{I}^N$ be an efficient, Hurwicz compatible solution in $\mathcal{IG}$. Let $iv \in \mathcal{IG}$. Let $\mathbf{1} \in [0, 1]^N$ defined as $\mathbf{1}_j = 1$ for all $j \in N$, and let $\mathbf{0} \in [0, 1]^N$ defined as $\mathbf{0}_j = 0$ for all $j \in N$. Then, for each $iv \in \mathcal{IG}$,

$$\overline{i\sigma_j(iv)} = \mathbf{1} \circ i\sigma_j(iv) = i\sigma_j(\mathbf{1} \circ iv) = i\sigma_j(\overline{iv}) \quad (8)$$

and

$$\underline{i\sigma_j(iv)} = \mathbf{0} \circ i\sigma_j(iv) = i\sigma_j(\mathbf{0} \circ iv) = i\sigma_j(\underline{iv}). \quad (9)$$

Since both $\overline{iv}$ and iv are TU games, we deduce that $i\sigma$ only depends on its restriction on TU games, i.e., once we define $i\sigma(v)$ for each $v \in \mathcal{G}$, we can deduce $i\sigma(iv)$ for any other $iv \in \mathcal{IG}$. Moreover, given $v, w \in \mathcal{G}$ with $v(S) \leq w(S)$ for all $S \subseteq N$, it holds $i\sigma(v) \leq i\sigma(w)$ (otherwise, $i\sigma(iv)$ with $iw(S) = [v(S), w(S)]$ for all $S \subseteq N$ would not be well-defined). Now, for each $v \in \mathcal{G}$, define $v^-, v^+ \in \mathcal{G}$ as follows:

$$v^+(S) := \begin{cases} v(N) & \text{if } v(S) \leq v(N) \\ v(S) & \text{if } v(S) > v(N) \end{cases}$$

$$v^-(S) := \begin{cases} v(S) & \text{if } v(S) \leq v(N) \\ v(N) & \text{if } v(S) > v(N) \end{cases}$$

for all $S \subseteq N$. Clearly, $v^-(S) \leq v(S) \leq v^+(S)$ for all $S \subseteq N$. Hence, $i\sigma(v^-) \leq i\sigma(v) \leq i\sigma(v^+)$. Moreover, $v^-(N) = v(N) = v^+(N)$ and hence $i\sigma(v^-) = i\sigma(v) = i\sigma(v^+)$. For each $x \in \mathbb{R}$, define $u^x \in \mathcal{G}$ as follows:

$$u^x(S) := \begin{cases} x & \text{if } S = N \\ 0 & \text{if } S \neq N \end{cases}$$

for all $S \subseteq N$. On one hand, either $u^{v(N)}(S) \leq v^+(S)$ for all $S \subseteq N$ (when $v(N) > 0$) or $v^-(S) \leq u^{v(N)}(S)$ for all $S \subseteq N$ (when $v(N) \leq 0$). On the other hand, $u^{v(N)}(N) = v(N) = v^+(N) = v^-(N)$. Hence, either $i\sigma(u^{v(N)}) = i\sigma(v^+)$ (when $v(N) > 0$) or $i\sigma(u^{v(N)}) = i\sigma(v^-)$ (when $v(N) \leq 0$). In either case,

$$i\sigma(v) = i\sigma(u^{v(N)}) \quad (10)$$

i.e. the only relevant value is $v(N)$. Let

$$f(x) := i\sigma(u^x) \quad (11)$$

for all $x \in \mathbb{R}$. Now, we prove that

$$f_j(\alpha_0 \cdot x) = \alpha_0 \cdot f_j(x) + (1 - \alpha_0) \cdot f_j(0) \quad (12)$$

for all $j \in N$, all $\alpha_0 \in [0, 1]$, and all $x \in \mathbb{R}$. We assume $x > 0$. Case $x < 0$ is analogous and case $x = 0$ is trivial. Define $iu^{0x} \in \mathcal{IG}$ as follows:

$$iu^{0x}(S) := \begin{cases} [0, x] & \text{if } S = N \\ 0 & \text{if } S \neq N \end{cases}$$

for all $S \subseteq N$. Let $\alpha \in [0, 1]^N$ defined by $\alpha_j = \alpha_0$ for all $j \in N$ for some $\alpha_0 \in [0, 1]$. Hence, for each $j \in N$,

$$\begin{aligned} f_j(\alpha_0 \cdot x) &= f_j(\alpha_0 \cdot x + (1 - \alpha_0) \cdot 0) = i\sigma_j(u^{\alpha_0 \cdot x + (1 - \alpha_0) \cdot 0}) \\ &= (i\sigma)_j(\alpha \circ iu^{0x}) = (i\sigma \circ \alpha)_j(iu^{0x}) \end{aligned}$$

by uniform Hurwicz compatibility of $i\sigma$:

$$\begin{aligned} &= (\alpha \circ i\sigma)_j(iu^{0x}) = \alpha_0 \circ i\sigma_j(iu^{0x}) \\ &= \alpha_0 \circ [i\sigma_j(iu^0), i\sigma_j(iu^x)] \\ &= \alpha_0 \cdot i\sigma_j(iu^x) + (1 - \alpha_0) \cdot i\sigma_j(iu^0) \\ &= \alpha_0 \cdot f_j(x) + (1 - \alpha_0) \cdot f_j(0) \end{aligned}$$

and hence (12) holds. This implies that there exist $\delta_j, \gamma_j \in \mathbb{R}$ such that

$$f_j(x) = \delta_j + \gamma_j \cdot x \quad (13)$$

for all $x \in \mathbb{R}$ (see Figure 1).

Clearly, $\delta_j = f_j(0)$. Moreover, $\gamma_j \geq 0$ because otherwise $i\sigma_j(iv) \notin \mathbb{I}$ when $|iv|(N) > 0$. By efficiency of $i\sigma$:

$$\sum_{j \in N} \delta_j = \sum_{j \in N} f_j(0) = \sum_{j \in N} i\sigma_j(u^0) = u^0(N) = 0$$

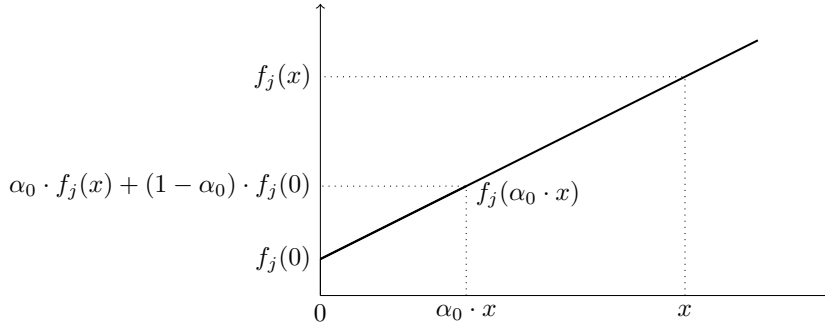


Figura 1: Visual proof that (12) implies (13).

and

$$\sum_{j \in N} \gamma_j = \sum_{j \in N} (\gamma_j \cdot 1) = \sum_{j \in N} (f_j(1) - \delta_j) = \sum_{j \in N} i\sigma_j(u^1) - \sum_{j \in N} \delta_j = u^1(N) - 0 = 1.$$

Finally, for all $iv \in \mathcal{IG}$ and all $j \in N$,

$$\begin{aligned} i\sigma_j(iv) &\stackrel{(9)(8)}{=} [i\sigma_j(\underline{iv}), i\sigma_j(\overline{iv})] \\ &\stackrel{(10)}{=} [i\sigma_j(u^{\underline{iv}(N)}), i\sigma_j(u^{\overline{iv}(N)})] \\ &\stackrel{(11)}{=} [f_j(\underline{iv}(N)), f_j(\overline{iv}(N))] \\ &\stackrel{(13)}{=} [\delta_j + \gamma_j \cdot \underline{iv}(N), \delta_j + \gamma_j \cdot \overline{iv}(N)] \\ &= \delta_j \cdot [1, 1] + \gamma_j \cdot [\underline{iv}(N), \overline{iv}(N)] = \delta_j \cdot [1, 1] + \gamma_j \cdot iv(N). \end{aligned}$$

□ ■

In particular, when $\delta_j = \delta_k$ and $\gamma_j = \gamma_k$ for all $j, k \in N$, then $\delta_j = 0$ and $\gamma_j = \frac{1}{n}$ for all $j \in N$, and so we obtain the *interval egalitarian* solution:

$$iE_j(iv) := \frac{1}{n} \cdot iv(N)$$

for all $j \in N$.

An important implication of Theorem 1 is the following:

Theorem 2 *The interval egalitarian solution is the only efficient, symmetric, and uniform Hurwicz compatible solution in $\mathcal{IG}$.*

Proof. It follows from Theorem 1 and the fact that the interval egalitarian solution is the only symmetric one in the family of efficient, uniform Hurwicz compatible interval solutions in $\mathcal{IG}$. □ ■

Corollary 2 *For $n > 1$, the interval HSX solution is not uniform Hurwicz compatible in $\mathcal{IG}$.*

Proof. It follows from Theorem 2 and the fact that the interval HSX solution is efficient, symmetric, defined in all $\mathcal{IG}$, and (for $n > 1$) different from the interval egalitarian solution in $\mathcal{IG}$. □ ■

Let $\mathcal{IC}$ denote the set of interval coalitional games $iv \in \mathcal{IG}$ satisfying $\underline{iv}(N) = \overline{iv}(N)$. Clearly, $\mathcal{G} \subset \mathcal{IC}$.

A simple family of efficient solutions in $\mathcal{IC}$ is given by

$$iE_j^\omega(iv) = \omega_j(iv(N))$$

for all $j \in N$, where $\omega(x) \in \mathbb{R}^N$ satisfies $\sum_{j \in N} \omega_j(x) = x$ for all $x \in \mathbb{R}$. In particular, when $\omega_j(x) = \omega_k(x)$ for all $j, k \in N$ and all $x \in \mathbb{R}$, we obtain the interval egalitarian solution restricted to $\mathcal{IC}$.

Theorem 3 An efficient solution $i\sigma : \mathcal{IC} \rightarrow \mathbb{R}^N$ is Hurwicz compatible in $\mathcal{IC}$ if and only if there exists a function ω that assigns to each $x \in \mathbb{R}$ a vector $\omega(x) \in \mathbb{R}^N$ with $\sum_{j \in N} \omega_j(x) = x$ and such that

$$i\sigma(iv) = iE^\omega(iv)$$

for all $iv \in \mathcal{IC}$.

Proof. Since we are in $\mathcal{IC}$, $i\sigma \circ \alpha = i\sigma$ for all efficient solution $i\sigma$ and all $\alpha \in [0, 1]^N$. Hence, Hurwicz compatibility is equivalent to

$$i\sigma(iv) = \sigma(\alpha \circ iv)$$

for all $iv \in \mathcal{IC}$ and all $\alpha \in [0, 1]^N$. It is not difficult to check that, given $\omega : \mathbb{R} \rightarrow \mathbb{R}^N$ with $\sum_{j \in N} \omega_j(x) = x$ for all $x \in \mathbb{R}$, iE^ω is Hurwicz compatible in $\mathcal{IC}$. Let $i\sigma : \mathcal{IC} \rightarrow \mathbb{R}^N$ be an efficient Hurwicz compatible solution for $\mathcal{IC}$. For each $x \in \mathbb{R}$, let $ie^x \in \mathcal{IC}$ defined as $ie^x(N) = [x, x]$ and $ie^x(S) = [0, 0]$ otherwise. Let

$$\omega_j(x) := i\sigma_j(ie^x)$$

for all $x \in \mathbb{R}$ and all $j \in N$. By efficiency of σ , we deduce $\sum_{j \in N} \omega_j(x) = x$ for all $x \in \mathbb{R}$. We prove that $\sigma(iv) = iE^\omega(iv)$ for all $iv \in \mathcal{IC}$ by induction on the cardinality of

$$\Theta(iv) = \{S \subset N, S \neq \emptyset : iv(S) \neq [0, 0]\}.$$

Assume first $|\Theta(iv)| = 0$. Then, $iv = ie^{iv(N)}$ and hence

$$i\sigma(iv) = i\sigma(ie^{iv(N)}) = \omega(iv(N)) = iE^\omega(iv).$$

Assume now the result holds when the cardinality of $\Theta(iv)$ is less than $\theta > 0$, and let $|\Theta(iv)| = \theta$. Let $S \in \Theta(iv)$. We have three cases:

Case 1: $0 \in iv(S)$. Since $S \in \Theta(iv)$, we deduce that there exists some $\alpha_0 \in [0, 1]$ such that $0 = \alpha_0 \circ iv(S)$.

Let $\alpha_j = \alpha_0$ for all $j \in N$ and let $iv^{-S} \in \mathcal{IC}$ defined as follows:

$$iv^{-S}(T) := \begin{cases} iv(T) & \text{if } T \neq S \\ [0, 0] & \text{if } T = S \end{cases}$$

for all $T \subseteq N$. It is straightforward to check that $\alpha \circ iv = \alpha \circ iv^{-S}$. Under Hurwicz compatibility of $i\sigma$ and the induction hypothesis,

$$\begin{aligned} i\sigma(iv) &= \sigma(\alpha \circ iv) = i\sigma(\alpha \circ iv^{-S}) = i\sigma(iv^{-S}) \\ &= iE^{\omega(iv^{-S}(N))}(iv^{-S}) = iE^{\omega(iv(N))}(iv). \end{aligned}$$

Case 2: $0 < \underline{iv}(S)$. Let $\alpha_j = 1$ for all $j \in N$. Let $iv^{-0,S} \in \mathcal{IC}$ defined as follows:

$$iv^{-0,S}(T) := \begin{cases} iv(T) & \text{if } T \neq S \\ [0, \overline{iv}(S)] & \text{if } T = S \end{cases}$$

for all $T \subseteq N$. It is straightforward to check that $\alpha \circ iv = \alpha \circ iv^{-0,S}$. Under Hurwicz compatibility of σ ,

$$\sigma(iv) = \sigma(\alpha \circ iv) = \sigma(\alpha \circ iv^{-0,S}) = \sigma(iv^{-0,S})$$

and we proceed as in Case 1 with $iv^{-0,S}$.

Case 3: $\overline{iv}(S) < 0$. Let $\alpha_j = 0$ for all $j \in N$. Let $iv^{+0,S} \in \mathcal{IC}$ defined as follows:

$$iv^{+0,S}(T) := \begin{cases} iv(T) & \text{if } T \neq S \\ [\underline{iv}(S), 0] & \text{if } T = S \end{cases}$$

for all $T \subseteq N$. It is straightforward to check that $\alpha \circ iv = \alpha \circ iv^{+0,S}$. Under Hurwicz compatibility of σ ,

$$\sigma(iv) = \sigma(\alpha \circ iv) = \sigma(\alpha \circ iv^{+0,S}) = \sigma(iv^{+0,S})$$

and we proceed as in Case 1 with $(iv^{+0,S})$.

□ ■

An important implication of Theorem 3 is the following:

Theorem 4 *The interval egalitarian solution is the only efficient and symmetric Hurwicz compatible solution in $\mathcal{IC}$.*

Proof. It follows from Theorem 3 and the fact that the interval egalitarian solution is the only symmetric one in the family of efficient, Hurwicz compatible interval solutions in $\mathcal{IC}$. □ ■

Let us consider a cooperative interval game $iv \in \mathcal{IG}$ and $\alpha \in [0, 1]^N$. For any coalition $S \subseteq N$, we denote $\alpha_S = \max_{j \in S} \alpha_j$. The core of the TU game $\alpha \circ iv \in \mathcal{G}$ defined as in (4) is the set:

$$C(\alpha \circ iv) := \left\{ x \in \mathbb{R}^N : \sum_{j \in N} x_j = \alpha_N \circ iv(N), \alpha_S \circ iv(S) \leq \sum_{j \in S} x_j, \forall S \subseteq N \right\}.$$

The interval core with N players as defined in [3] is the set

$$iC(iv) := \left\{ ia \in \mathbb{I}^N : \sum_{j \in N} ia_j = iv(N), iv(S) \preccurlyeq \sum_{j \in S} ia_j, \forall S \subseteq N \right\}.$$

In this case, we can apply the Hurwicz criterion to $iC(iv)$ and obtain the set

$$\begin{aligned} (\alpha \circ iC)(iv) &= \alpha \circ iC(iv) \\ &= \{(\alpha \circ ia)_j \in \mathbb{R}^N : ia \in iC(iv)\} \\ &= \left\{ (\underline{ia}_j + \alpha_j |ia_j|)_{j \in N} \in \mathbb{R}^N : \sum_{j \in N} ia_j = iv(N), iv(S) \preccurlyeq \sum_{j \in S} ia_j, \forall S \subseteq N \right\}. \end{aligned}$$

It is clear that

$$\sum_{j \in N} ia_j = iv(N) \iff \sum_{j \in N} \underline{ia}_j = \underline{iv}(N) \text{ and } \sum_{j \in N} |ia_j| = |iv|(N)$$

and for any coalition $S \subseteq N$ we have

$$iv(S) \preccurlyeq \sum_{j \in S} ia_j \iff \underline{iv}(S) \leq \sum_{j \in S} \underline{ia}_j \text{ and } |iv|(S) \leq \sum_{j \in S} |ia_j|.$$

Proposition 4 *If all the Hurwicz coefficients coincide, $\alpha_1 = \alpha_2 = \dots = \alpha_n = \alpha_0$, then*

$$\alpha \circ iC(iv) \subseteq C(\alpha \circ iv). \quad (14)$$

Proof. Let us consider $x \in \alpha \circ iC(iv)$. Then, $x_j = \underline{ia}_j + \alpha_0 |ia_j|$ for each $j \in N$ and some $ia \in \mathbb{I}^N$ with $\sum_{j \in N} \underline{ia}_j = \underline{iv}(N)$, $\sum_{j \in N} |ia_j| = |iv|(N)$, $\underline{iv}(S) \leq \sum_{j \in S} \underline{ia}_j$ and $|iv|(S) \leq \sum_{j \in S} |ia_j|$ for all $S \subseteq N$. Then,

$$\sum_{j \in N} x_j = \sum_{j \in N} \underline{ia}_j + \alpha_0 \sum_{j \in N} |ia_j| = \alpha_0 \circ iv(N)$$

and, for any $S \subseteq N$,

$$\alpha_0 \circ iv(S) \leq \sum_{j \in S} x_j = \sum_{j \in S} \underline{ia}_j + \alpha_0 \sum_{j \in S} |ia_j|$$

so $x \in C(\alpha \circ iv)$. □ ■

Observe that inclusion (14) also holds when $iC(iv) = \emptyset$. Inclusion $C(\alpha \circ iv) \subseteq \alpha \circ iC(iv)$ is not always true, as in the following example.

Example 1 *Consider $n = 2$ and the interval cooperative game iv given by $iv(\{1\}) = iv(\{2\}) = [1, 3]$ and $iv(\{1, 2\}) = [2, 4]$. For this game, $iC(iv) = \emptyset$, while if we consider $\alpha_1 = \alpha_2 = 0$, game $\alpha \circ iv$ has non empty core: $0 \circ iv(\{1\}) = 1 \circ iv(\{2\}) = 1$ and $0 \circ iv(\{1, 2\}) = 2$, so $C(\alpha \circ iv) = \{(1, 1)\}$.*

Then we can deduce the next result.

Proposition 5 *The interval core is not uniform Hurwicz compatible in $\mathcal{IG}$.*

Remark 1 *In we consider in the previous example certainty in the grand coalition, in particular if we assume $iv(\{1, 2\}) = [2, 2]$, we have still that the inclusion $C(\alpha \circ iv) \subseteq \alpha \circ iC(iv)$ does not hold.*

Given an interval cooperative game $iv \in \mathcal{IG}$, we call *border games* the two (classical) cooperative games defined as $\underline{iv}$ and $\overline{iv}$ where for any $S \subseteq N$, $iv(S) = [\underline{iv}(S), \overline{iv}(S)]$. In [5] the *square interval core* $C(\underline{iv}) \square C(\overline{iv})$ is defined as

$$C(\underline{iv}) \square C(\overline{iv}) := \left\{ ia \in \mathbb{I}^N : \sum_{j \in N} ia_j = iv(N), \underline{ia} \in C(\underline{iv}), \overline{ia} \in C(\overline{iv}) \right\}$$

and it has been proved that if $iC(iv) \neq \emptyset$ then $C(\underline{iv}) \square C(\overline{iv}) = iC(iv)$. Then, the same considerations of the interval core on the Hurwicz compatibility can be done.

4. CONCLUDING REMARKS

In this paper we study interval cooperative games. These are games where the worth of coalitions are uncertain. Both a lower and an upper bound of the possible final outcome is assigned to each coalition. For these games, several solution concepts provide interval allocations to the players and leave uncertainty on the exit. In order to mitigate this uncertainty, assuming some degree of optimism (or pessimism) of the players (given by real numbers between 0 and 1), we introduce a TU cooperative game applying the Hurwicz criterion. This procedure allows to have a standard solution concept once the degree of optimism is fixed.

Another possibility for approaching the uncertainty is the following: consider any interval solution concept in the original interval game and then apply the Hurwicz criterion to the interval allocation. The question posed in the paper is if the two approaches lead to the same result. The idea of Hurwicz compatibility is given and we investigate under which conditions it holds. In the case of a uniform degree of optimism/pessimism or of the grand coalition certainty, we prove that the only compatible solutions are the proportional ones, or the egalitarian in case symmetry is required. Some considerations on the Shapley value and the core solution are also discussed.

REFERENCES

- [1] Abe T. and Nakada S. The weighted-egalitarian Shapley values. *Social Choice and Welfare* 52 (2), pp. 197-213 (2019).
- [2] Alparslan Gök, S., Brânzei, R., and Tijs, S., Cores and stable sets for interval-valued games. Discussion Paper number 2008-17 2008-17, Tilburg University, Center for Economic Research (2008).
- [3] Alparslan Gök, S., Brânzei, R., and Tijs, S., Airport interval games and their Shapley value. *Operations Research and Decision*, vol. 2, pp. 571-575 (2009a).
- [4] Alparslan Gök, S., Brânzei, R., and Tijs, S., Convex interval games. *Journal of Applied Mathematics and Decision Sciences* vol. 2009 (14) Article ID 342089 (2009b).
- [5] Alparslan Gök, S., Miquel, S., and Tijs, S., Cooperation under interval uncertainty. *Mathematical Methods of Operations Research* vol. 69, pp. 99-109 (2009c).
- [6] Alparslan Gök, S. Z., Brânzei, R., Fragnelli, V., and Tijs, S., Sequencing interval situations and related games. *Central European Journal of Operations Research* vol. 21, pp. 225-236 (2013).
- [7] Alparslan Gök, S. Z., Brânzei, R., and Tijs, S. The interval Shapley value: an axiomatization. *Central European Journal of Operations Research*, vol. 18 (2), pp. 131-140 (2010).
- [8] Béal, S., Casajus, A., Huettner, F., Rémila, E., and Solal, P. Characterizations of weighted and equal division values. *Theory and Decision*, vol. 80 (4), pp. 649-667 (2016).

- [9] Béal, S., Rémila, E., and Solal, P. Coalitional desirability and the equal division value. *Theory and Decision* 86 (1), pp. 95-106 (2019).
- [10] Bergantiños, G., and Vidal-Puga, J. Additive rules in bankruptcy problems and other related problems. *Mathematical Social Sciences* 47 (1), pp. 87-101 (2004).
- [11] Bergantiños, G., and Vidal-Puga, J. A value for PERT problems. *International Game Theory Review* 11 (4), pp. 419-436 (2009).
- [12] Bolton, G.E., and Ockenfels, A. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* 90 (1), pp. 166-193 (2000).
- [13] Brânzei, R. and Alparslan Gök, S. Z., Bankruptcy problems with interval uncertainty. *Economics Bulletin*, vol. 3 (56), pp. 1-10 (2008).
- [14] Brânzei, R., Brânzei, O., Alparslan Gök, S. Z., and Tijs, S. Cooperative interval games: a survey. *Central European Journal of Operations Research*, vol. 18 (3), pp. 397-411 (2010).
- [15] Casajus, A. and Huettnner, F. Null players, solidarity, and the egalitarian Shapley values. *Journal of Mathematical Economics*, vol. 49 (1), pp. 58-61 (2013).
- [16] Casajus, A. and Huettnner, F. Weakly monotonic solutions for cooperative games. *Journal of Economic Theory*, vol. 154, pp. 162-172 (2014).
- [17] Casajus A. and Huettnner F. On a class of solidarity values. *European Journal of Operational Research* 236 (2), pp. 583-591 (2014).
- [18] Dutta, B. and Ray, D. A concept of egalitarianism under participation constraints. *Econometrica*, vol. 57, pp. 615-535 (1989).
- [19] Han, W., Sun, H., and Xu, G. A new approach of cooperative interval games: The interval core and Shapley value revisited. *Operations Research Letters*, vol. 40 (6), pp. 462-468 (2012).
- [20] Hurwicz, L. The generalized Bayes minimax principle: a criterion for decision making under uncertainty. Discussion paper 335, Cowles Commission (1951).
- [21] Hougaard, J. L. and Moulin, H. Sharing the cost of risky projects. *Economic Theory*, vol. 65 (3), 663-679 (2018).
- [22] Koster, M. and Boonen, T. J. Constrained Stochastic Cost Allocation. Working paper at University of Amsterdam (2019).
- [23] Lardon, A. Endogenous interval games in oligopolies and the cores. *Annals of Operations Research*, vol. 248 (1), pp. 345-363 (2017).
- [24] Li, D.-F. *Models and Methods for Interval-Valued Cooperative Games in Economic Management*. Springer (2016).
- [25] Montemanni, R. A Benders decomposition approach for the robust spanning tree problem with interval data. *European Journal of Operational Research*, vol. 174 (3), pp. 1479-1490 (2006).
- [26] Moretti, S., Gök, S. Z. A., Brânzei, R., and Tijs, S. Connection situations under uncertainty and cost monotonic solutions. *Computers and Operations Research*, vol. 38 (11), pp. 1638-1645 (2011).
- [27] Pereira, J. and Averbakh, I. Exact and heuristic algorithms for the interval data robust assignment problem. *Computers and Operations Research*, vol. 38 (8), pp. 1153-1163 (2011).
- [28] Shapley, L. S. A value for n-person games. In Kuhn, H. and Tucker, A., editors, *Contributions to the theory of games*, volume II of *Annals of Mathematics Studies*, pp. 307-317. Princeton University Press, Princeton NJ. (1953).
- [29] Thrall, R. and Lucas, W. N-person games in partition function form. *Naval Research Logistic Quarterly*, vol. 10, pp. 281-298 (1963).

-
- [30] van den Brink, R. Null or nullifying players: The difference between the Shapley value and equal division solutions. *Journal of Economic Theory*, vol. 136, pp. 767-775 (2007).
 - [31] van den Brink, R., Chun, Y., Funaki, Y., and Park, B. Consistency, population solidarity, and egalitarian solutions for TU-games. *Theory and Decision*, vol. 81 (3), pp. 427-447 (2015).
 - [32] van den Brink, R., Palanci, O., and Gök, S. Z. A. Interval solutions for TU-games. Ti 2017-094/ii, Tinbergen Institute (2017).
 - [33] Wu, W., Iori, M., Martello, S., and Yagiura, M. Exact and heuristic algorithms for the interval min-max regret generalized assignment problem. *Computers and Industrial Engineering*, vol. 125, pp. 98-110 (2018).
 - [34] Xue, J. Fair division with uncertain needs. *Social Choice and Welfare*, vol. 51 (1), pp. 105-136 (2018).
 - [35] Yokote, K., Kongo, T., and Funaki, Y. The balanced contributions property for equal contributors. *Games and Economic Behavior*, vol. 108, pp. 113-124. Special Issue in Honor of Lloyd Shapley: Seven Topics in Game Theory (2018).

OS TABULADORES, AS SÚAS COMPOÑENTES E MAIS A SÚA UTILIDADE

Carlos L. Iglesias Patiño¹

¹ Instituto Galego de Estatística

RESUMO

Introducimos funcións de conxunto ou argumento nominal que serven para tabular con diferentes finalidades: describir unha poboación finita, construír indicadores ou mesmo calcular estimativas. Estas aplicacións tabuladoras baséanse no criterio dos cadrados mínimos e tratan de formalizar e criticar a ubicua proporcionalidade na estatística pública.

Palabras e frases chave: tabulador, descriptor, indicador, suavización local, factor de separación

1. INTRODUCCIÓN

Existe unha gran demanda de información estatística, parcialmente insatisfeita. Estamos perante unha exuberancia de datos aínda que non se dispón dunha oferta de información estatística suficiente. Para tratar este desequilibrio, a estatística pública está a vivir tempos de mudanza tanto na captación dos datos como na difusión dos resultados estatísticos.

Antano existía unha diferenza nidia entre datos estatísticos e datos de orixe administrativa ou doutra índole xurídica, entre estatísticas exhaustivas e por mostraxe ou entre información estatística e cartográfica. Hogano os datos de orixe non estatística posúen un interese cada vez meirande inda máis, se cadra, con escenarios orzamentarios restritivos. Na actualidade é máis frecuente que se mesturen este tipo de datos cos de carácter estatístico para elaborar, á súa vez, unha nova estatística (López e Iglesias, 2017; IGE, 2018 e mais Torrado e López, 2017).

Un exemplo paradigmático foron os censos demográficos de 2011 para os que existía unha base de recenseamento, o Padrón Municipal de Habitantes, que funciona en España como rexistro continuo de poboación. Existen outras situacións nas que non predomina ningunha fonte como cando distintos órganos ou, mesmo, diferentes administracións públicas ostentan a mesma competencia sobre diferentes zonas dun espazo xeográfico, por exemplo, nas estatísticas da auga da Comunidade Autónoma (C.A.) de Galicia conviven dúas administracións públicas diferentes, Estado e C.A.; dentro do Estado, tres organismos de bacía (Cantábrica, Miño-Sil e Douro).

Neste caso, salvante as distancias, tamén se atoparía a estatística europea en xeral que centraliza e coordena información que non sempre é de idéntica natureza inda que estea harmonizada. Isto comporta certas consecuencias como, por exemplo, igual actuación administrativa mais cun calendario lixeiramente diferente, convivencia de datos observados con datos estimados ou imputados. Ambas as situacións entrañan o que Azorín (1986) denominaba o problema da rexionalización, interna e externa respectivamente. Como resumo, chégase a dispor dos micro-datos necesarios mais con meta-datos diferentes abondo.

Outro tipo de información estatística amplamente demandada corresponde máis ao cumio da pirámide que á súa base, é dicir, resultados máis elaborados que inclúen a presentación non convencional de operacións básicas (Teijeiro, 1991), os sistemas de indicadores, a información espacial e sectorial a un nivel máis desagregado do proxectado inicialmente ou os balances en unidades físicas.

Dentro dos indicadores predominaron, ata agora, os que era doado presentar no formato tabular. Con todo xa é hora de afrontar que un sistema de indicadores non pode estar condicionado fundamentalmente pola forma. Debe liberarse do ‘corset’ do formato, isto é, o deseño dun indicador debe ser, á partida, independente dos medios de difusión. Debemos avanzar no emprego de indicadores funcionais ou de campo, no senso xeo-estadístico, e como mínimo nos vectoriais.

Os indicadores funcionais xunto coa súa representación gráfica, ademais de achegar claridade e riqueza, han redundar nunha mellor validación externa e na macro-depuración dos datos estatísticos.

Unha construción xeométrica achega unha cantidade de detalles superior a un número ou a un vector aínda que teña menos precisión e resulte menos sumaria (Iglesias e López, 2018).

Todos estes cambios recomentan variacións nos métodos estatísticos empregados. A multiplicidade de fontes constitúe un novo xerador de variabilidade que pode aconsellar o emprego do alisamento ou suavización. A súa necesidade xa se manifestou no uso da información estatística en contextos como as análises demográfica e da conxuntura ou nas estatísticas de síntese (contas, balances...), desta hora maniféstase en operacións estatísticas de carácter básico.

Con todo, sen sermos tan ambiciosos e conformándonos con resultados intermedios, conviría tratar de avanzarmos incardinando situacións usuais como as distribucións xeográficas e os seus cartogramas, campos incipientes, ou as distribucións segundo un nivel de desagregación dunha clasificación estándar (CNAE, CNO, CNED...) dentro dun cadro conceptual que permita nun futuro achegarse a eses ‘desiderata’.

En numerosos casos pola natureza do problema, como os exemplos anteriores, ou por necesidades da difusión non se poderá avanzar e estes ‘desiderata’ serán inatinxíbeis porque o conxunto inicial nesas situacións non é numérico inda que o seu conxunto final sexa $\mathbb{R}^n$. Non sempre se está na presenza dunha variábel senón dun carácter, en principio, cualitativo.

Podemos estar interesados en estudar unha variábel Y que pode estar influída por dous caracteres, un cualitativo e outro cuantitativo. O primeiro ha funcionar como un carácter de clasificación e o segundo como unha variábel complementaria. Por iso, convén introducir unhas funcións de conxunto, ou por brevidade simplemente aplicacións, que denominaremos tabuladores.

2. TABULADORES

Por todo o anterior, convén xeneralizar os conceptos usuais da estatística pública. Sexa U un colectivo con I unidades estatísticas. De cada unidade i , estudamos dúas variábeis X e Y e mais un carácter Z que induce una partición $\Pi = \{B_c\}_{c=1}^C$ de U , sendo C o número de clases nas que dividimos o colectivo, que podemos denotar $\{(i, z_i, X_i, Y_i), i = 1, \dots, I\}$. A Y é a variábel principal mentres a X e o Z han servir para estudar aquela dun xeito mellor.

Exemplos paradigmáticos de Z son un carácter nominal no que empregamos unha clasificación hierárquica, a cada nivel da clasificación dispónse dunha partición Π_v por exemplo en Galicia, entidade singular, entidade colectiva, municipio, comarca, área funcional, provincia, C.A., E.M., –en xeral, NUTS de 5 a 1–, Unión Monetaria, Unión Europea; ou ben o dun carácter cuantitativo, variábel Z , que inda que tome valores distintos nun intervalo, discretizámola mediante subintervalos de igual lonxitude, h_n , por tanto para cada n terase unha partición diferente.

A partición correspondente ao carácter de clasificación pode estar determinada por estándares, por tradición ou vir condicionada polo segredo estatístico. Exemplos do primeiro tipo eran a “Recomendación da Comisión de 6 de maio de 2003 sobre a definición de micro-empresas, pequenas e medianas empresas” e o “Regulamento (UE) 275/2010 da Comisión de 30 de marzo de 2010 polo que se aplica o Regulamento (CE) 295/2008 do Parlamento Europeo e do Consello, no relativo aos criterios de avaliación da calidade das estatísticas estruturais das empresas”, en proceso de reforma ou mesmo xa derrogadas.

Exemplos da complementaria X son unha contaxe (variábel discreta) ou outra variábel de tamaño (continua), de feito esta podémola ver como unha xeneralización da variábel de conta. Tamén pode tentar medir a presenza ante o proceso de observación ou ben a ponderación de estarmos perante un colectivo de unidades heteroxéneas.

Dentro do paradigma da proporcionalidade, adoito empregado na estatística pública e noutros campos mesmo na aprendizaxe automática –onde algúns autores cando tratan de regresión linear ou de redes neurais falan de parámetro de nesgo referíndose á resposta na orixe–, mais de xeito local, podemos definir unhas aplicacións con orixe na partición xerada por Z que operan sobre o colectivo, en concreto, introducimos os tabuladores que resultan do seguinte criterio

$$R_g(B_c) = \arg \min_R \sum_{i=1}^I (Y_i - RX_i)^2 X_i^{-g} 1_{B_c}(i),$$

nomeadamente, habemos tratar $g = 0, 1$ e 2 . A ecuación normal é

$$\sum_{i=1}^I (Y_i - R_g(B_c)X_i) X_i^{1-g} 1_{B_c}(i) = 0$$

da que resulta que

$$R_g(B_c) = \frac{\sum_{i=1}^I X_i^{1-g} Y_i 1_{B_c}(i)}{\sum_{i=1}^I X_i^{2-g} 1_{B_c}(i)}$$

No caso da tabulación no senso de Gz.Duaso (2018), isto é, unha agregación ‘metadata driven’, existe un problema de definición cando non existen unidades que pertencen á clase. Nese caso a ecuación normal verificase trivialmente con calquera R . Optaremos por definir $R(B_c) := 0$ neste caso patolóxico, por resultar práctico como se amosará posteriormente. Pode tomarse outra decisión como prescindir destas clases por considerarmos que non proceden, semellante á condensación en terminoloxía deste mesmo autor, agregación ‘data driven’. Nas clases non-baleiras, a ecuación normal e o tabulador son respectivamente:

$$\sum_{B_c} (Y_i X_i^{1-g} - R(B_c) X_i^{2-g}) = 0$$

$$R_g(B_c) = \frac{\sum_{B_c} X_i^{1-g} Y_i}{\sum_{B_c} X_i^{2-g}}$$

Para $g = 0$, a ecuación normal redúcese a $\sum_{i=1}^I (Y_i - R(B_c) X_i) X_i 1_{B_c}(i) = 0$, que consiste nun axustamento linear local pola orixe, e nas celas non-baleiras, $\sum_{B_c} X_i Y_i - \sum_{B_c} X_i R(B_c) X_i = 0$, e mais

$$R_0(B_c) = \frac{\sum_{B_c} X_i Y_i}{\sum_{B_c} X_i^2}$$

que podémolo interpretar como o cociente entre o total local da variábel produto e o do cadrado da complementaria ou tamén o momento produto local ($A_{11}(B_c)$) dividido polo momento local de orde dúas da complementaria ($A_{20}(B_c)$).

Se $g = 1$, a ecuación normal é $\sum_{i=1}^I (Y_i - R(B_c) X_i) 1_{B_c}(i) = 0$, e nas celas non-baleiras

$$\sum_{B_c} Y_i - R(B_c) \sum_{B_c} X_i = 0,$$

$$R_1(B_c) = \frac{\sum_{B_c} Y_i}{\sum_{B_c} X_i}.$$

Interpretámolo como o cociente entre os totais locais da variábel Y e da X ou tamén como o cociente das medias aritméticas locais, $A_{01}(B_c)/A_{10}(B_c)$. Observamos que, neste caso, o total local de Y coincide co produto do total local de X polo tabulador. Isto é, podemos construír un novo colectivo aproximado $\{(i, z_i, X_i, \hat{Y}_i), i = 1, \dots, I\}$ mediante o tabulador, $\hat{Y}_i = \sum_{c=1}^C R_1(B_c) X_i 1_{B_c}(i)$. Este novo colectivo caracterízase por ter o mesmo total que o orixinal, por iso en estatística pública habémolo considerar como o tabulador por antonomasia.

Á derradeira, para $g = 2$ nas celas non-baleiras, $\sum_{B_c} (Y_i X_i^{-1} - R(B_c)) = 0$ e $R_2(B_c)$ é a media aritmética local das razóns ou na notación de momentos locais, $A_{-1,1}(B_c)/A_{00}(B_c)$.

Este esquema, que dende un punto de vista matemático resulta un pouco restritivo, permite formalizar múltiples tarefas que se executan adoito na estatística pública como amosamos a seguir sen pretender sermos exhaustivos. Por exemplo, estamos a tratar con verdadeiras unidades elementais e o colectivo é un universo que pretendemos estudar. Estes tabuladores poden ser interesantes para depurar estes micro-datos que constitúen a poboación finita (p.f.) e mais para explotala (meso-datos) e obter resultados estatísticos (macro-datos) para difundir como habemos estudar na seguinte subsección. Ou ben para deseñar un posíbel plan de mostraxe futuro (estratificación, conglomeración...).

Tamén podemos estar interesados en agregar polo carácter Z cando pretendemos difundir, trala debida depuración, resultados para unha poboación como a de concellos de Galicia ou para obtermos meso-datos que sirvan como insumos de operacións de síntese (táboas demográficas, matrices auxiliares do sistema de contas...). Mesmo podemos ir máis aló, considerar como colectivo unha destas poboacións xenéricas, formado por unidades estatísticas complexas (non elementais), e pretender estudármolo con variábeis procedentes de fontes diferentes. Isto habémolo estudar na subsección 2.2.

Á derradeira, o colectivo tamén pode ser unha mostra, aleatoria ou non, por ser máis complexa se é probabilística ou por ser, a priori, non-probabilística. Neste caso ademais de depuralo tamén podemos estar interesados en estimar ou imputar nun senso amplo. Na subsección 2.3, trataremos esta posibilidade.

2.1. DESCRITORES APLICATIVOS

Nesta subsección o colectivo U é un universo, formado por unidades elementais, cuxo cardinal adoita denotarse por N e a súa p.f. asociada, $\mathcal{P} = \{(i, z_i, X_i, Y_i): i = 1, \dots, N\}$. Se Z induce unha partición

con N subconxuntos unitarios, de ser cualitativo, pode resultar redundante, $\mathcal{P} = \{(i, X_i, Y_i), i = 1, \dots, N\}$. Outrosí ocorre no caso de Z ser constante, a inducida había ser $\{U, \emptyset\}$. Cando estamos preocupados por unha clase, estaríase en presenza da partición máis simple, á parte da trivial, a composta por un dominio e o seu co-dominio, o complementario del, $B_1 = D$ e $B_2 = CD$.

Habemos tentar xeneralizar os conceptos usuais da estatística descritiva –descriptores dunha poboación– a unha situación que abrangue mesmo os tradicionais que non couberon nun tratamento unitario como a tabulación, en senso amplo, e a representación gráfica. Para iso, habemos introducir o termo descriptores aplicativos. Estas aplicacións empregaranse para describir a p.f., por iso, recibirán este adxectivo que xeneraliza o de funcional.

A máxima ou aforismo “a mellor táboa dunha liña é unha frase ou unha alínea”, podémola mudar por a mellor táboa dunha liña é unha aplicación ou a súa representación. Habiamos denominalos descriptores funcionais, sempre que tivermos en mente unha función de conxunto, e abreviar co acrónimo DesFunCon (descriptor función de conxunto). En principio, estaríamos perante unha función real de argumento nominal, inda que preferimos escoller aplicación por brevidade reservándonos o termo función para argumentos numéricos.

De denominarmos subtotalizador á aplicación que asigna a cada clase da partición o seu total da variábel, $Y_+(B_c) := \sum_{i=1}^N Y_i 1_{B_c}(i)$, o denominador é o subtotalizador de X se $g = 1$. Para $g = 2$ é o recontador, isto é, a aplicación que asigna a cada clase a súa frecuencia absoluta, que habemos denotar $N(B_c)$. O tabulador é o cociente dos subtotalizadores de Y e de X . Nos outros tabuladores aparecen subtotalizadores de variábeis varias.

Cando consideramos a covariábel máis sinxela $X = 1$, os tabuladores ofrecen as medias locais de Y , $R_g(B_c) = \bar{Y}(B_c)$. Se ademais Y tamén é a constante unidade, $Y_i = X_i = 1$, entón ambos os subtotalizadores convértense no recontador e os tabuladores resultan

$$R_g(B_c) = \frac{N(B_c)}{1 \vee N(B_c)}$$

isto é, dannos se a clase está baleira ($R_g(B_c) = 0$) ou se dispón de unidades ($R_g(B_c) = 1$). Se Y é a indicatriz dunha outra característica entón o tabulador R_1 coincide coa fracción desa característica en B_c , a proporción.

Manténdomos $Y_i = 1$, unha elección artificial de X , tal que valla 1 na primeira unidade dos elementos da clase e 0 no resto, converte o R_1 no recontador. Se consideramos que X valla $N l(B_c) N^{-1}(B_c)$, no canto de 1, con $N = \sum_c N(B_c)$ e $l(B_c)$ a lonxitude do banzo, $[L_{c-1}, L_c]$ ou do chanzo $[L_{c-1}, L_c]$, o tabulador convertese no histograma que se pode estender considerándonos unha medida máis xeral $\mu(B_c)$, por exemplo a superficie, no canto de $l(B_c)$.

2.2. INDICADORES

Como resultado do proceso da subsección anterior, realizado por nós ou por outras persoas ou institucións e aplicado a unha ou a diversas fontes, podemos obter uns meso ou macro-datos referidos a unhas novas poboacións de unidades complexas, mesmo artificiais, como divisións administrativas, sectores, ramos de actividade, xénero e grupos etarios...

Nesta situación X pode ser unha variábel que permita relativizar –por exemplo a poboación permite a capitación– ou converter unidades ou, inclusive, construír o conversor, isto é, a transformación a valores monetarios comparábeis ou a unha mesma unidade física. Tamén pode formar parte dun proceso máis xeral de estandarización.

Uns dos labores dun órgano central das estatísticas dunha C.A. é a elaboración de compendios e sistemas de indicadores. Estes adoitan ser tabuladores, ou operacións aritméticas deles, definidos no mesmo conxunto orixe, o colectivo, inda que X , Y e Z poidan proceder de distinta fonte.

2.3. “ESTIMADORES”

Noutras situacións o colectivo é unha parte do universo. Nun caso, este subconxunto é resultado do deseño dunha mostra e o posterior procedemento de obtención efectiva dela non se afasta moito do deseñado. Noutros casos a mostraxe é empírica, non-probabilística (opinática, intencional ou por cotas) ou a obtención efectiva arredouse do deseño ou ben o colectivo procede dunha operación, á partida, exhaustiva mais que remata por ser parcial e cómpre dar resultados para toda a poboación inda que haxa que tomalos coa debida cautela dende unha perspectiva estatístico-matemática. Se pretendemos

aproximar un descriptor e a mostra é probabilística estamos perante un verdadeiro estimador, noutros casos, non sempre, de aquí as aspas do título da subsección.

Na situación de mostra probabilística inicial, Y sería a variábel de interese e X a probabilidade de inclusión ou de selección ou o inverso dos ponderadores tras sofícarlos (calibración), ou outro tipo de endereitamento da mostra. Isto é, en certo modo, estes ponderadores inversos constitúen unha medida da singularidade ou auto-representación da unidade. O ponderador semella o número de (co-)representados da clase.

Noutras situacións poderíamos pensar en tentar realizar unha estimación so modelo. Non temos avanzado moito neste eido dende que Iglesias (2004) avogase por alargar o emprego de modelos na mostraxe en estatística pública. A inferencia baseada no modelo de predición inda é case que inexistente, prefíren un tratamento sen especificaren o modelo. Neste procedemento heurístico desempeñan ou poderían desempeñar o seu papel os tabuladores, outro tanto na estimación baixo determinados modelos. Convén non esquecer que os descritores aplicativos poden empregarse, a súa vez, como estimadores de parámetros dunha superpoboación.

3. UTILIDADES

Ademais da utilidade na exploración dunha operación exhaustiva con omisos ‘missings’, na estratificación a priori ou a posteriori, equilibrio ou elevación das mostras, estas aplicacións podémolas empregar no cálculo de coeficientes técnicos ou operativos ou ben factores de conversión ou corrección de unidades. A seguir presentamos como exemplificación o tratamento do problema dos centros ou marcos de clase.

Este é un caso típico de meso-datos, non teñen como finalidade a difusión inda que poidan ser publicados sen grandes problemas. Abusando da linguaxe, o factor de separación é o incremento que debemos engadir ao extremo inferior dun intervalo de idade para obter o seu representante de clase. Este último semella a media de anos vividos polas cabezas que falecen no chanzo na táboa de mortalidade.

Podemos considerar como marco etario sinxelamente o centro da clase, neste caso o factor había ser a semilonxitude do intervalo $l(B_c)/2$. Ou ben empregar como factor de separación cifras estándar ofrecidas por outras institucións ou utilizar a información dispoñible de anos anteriores para afinar máis os resultados das esperanzas de vida (Martín et al., 2013 e IGE, 2017). Na terceira opción, de considerarmos idades simples, anos cumpridos, o factor de separación é a media local dos días vividos expresado en anos, neste caso un regresograma.

Cando pasamos a unha táboa de mortalidade abreviada, sexa B_c un semi-segmento etario —coincide cos quinquenios $B_c = [5c, 5(c+1))$ a partir de $c = 1$, en notación demográfica ${}_5B_{5c}$, agás a semi-recta derradeira ou chanzo superior ilimitado—. O primeiro, $c = 0$, escíndese entre menores dun ano e ${}_4B_1 = [1,5)$ —. O factor de separación convértese en $a_c = R_1(B_c; A_r) + R_1(B_c; A_c)$ onde

$$R_1(B_c; A_r) = \frac{\sum_{i=5c}^{5(c+1)-1} A_{r,i}}{\sum_{i=5c}^{5(c+1)-1} D_i} = \frac{\sum_{i=5c}^{5c+4} A_{r,i}}{\sum_{i=5c}^{5c+4} D_i}$$

con $A_{r,i}$ os días vividos no derradeiro ano de vida expresados en anos, anos residuais, e D_i as defuncións, e mais

$$R_1(B_c; A_c) = \frac{\sum_{i=5c}^{5c+4} (i - 5c) D_i}{\sum_{i=5c}^{5c+4} D_i},$$

A_c denota anos cumpridos. Isto é o factor de separación é a suma de dous tabuladores, o dos anos residuais e outro, o dos anos cumpridos no quinquenio, do tipo media ponderada.

No caso do primeiro cuadrenio,

$$R_1({}_4B_1; A_r) = \frac{\sum_{i=1}^4 A_{r,i}}{\sum_{i=1}^4 D_i}$$

e

$$R_1({}_4B_1; A_c) = \frac{\sum_{i=1}^4 (i - 1) D_i}{\sum_{i=1}^4 D_i}$$

e, dun xeito análogo, trataríamos as táboas abreviadas decenais.

4. CONCLUSIÓN

Este relatorio pretende continuar a anovar a estatística descriptiva para que converxa coa análise exploratoria de datos e que, como mínimo, se produza un aliñamento entre elas, tentando expresar

matematicamente algunhas das prácticas tradicionais nunha oficina das estatísticas. A introdución dos tabuladores pode facilitar o tratamento informático favorecendo a programación funcional, por exemplo, o subtotalizador pode escribirse empregándonos a notación de Iverson como $Y_+(B_c) = \sum_i Y_i[i \in B_c]$. Tamén amosa con máis claridade que estas prácticas constitúen unha sorte de suavización.

Unha codificación co maior detalle xunto coa presentación a un nivel menos pormenorizado dos valores resumo ou sintéticos semella un alisamento des dunha perspectiva matemática. O proceso de codificación e presentación de valores medios a diferentes niveis dunha clasificación hierárquica, como resultados cada vez máis agregados, constitúe unha suavización progresiva podendo chegar nas xeonomenclaturas ata a sobre-suavización de ofrecer o dato para Galicia ou nas clasificacións de actividades ata dar o dato para toda a economía galega ou os catro grandes sectores halo-agrario, construción, industria e servizos; en palabras de Leguina (1981): «[...] os índices globais ocultan máis que ensinan». O proceso descendente –cara ás follas na árbore hierárquica– de disgregación revela máis detalles así e todo tamén pode conducir a unha infra-suavización non difundíbel ou difícil de interpretar.

Por outra banda, esta contribución procura tamén unha reflexión máis profunda sobre a proporcionalidade que evite o emprego inadecuado dela na estatística pública. É certo que cando ten senso, inda con ser de xeito aproximado, resulta evidente a súa comodidade mais debe evitarse o seu uso indiscriminado porque se $Y_i \propto X_i$, os tres descritores $R_g(B_c)$, $g = 1,2,3$ coinciden coa constante de proporcionalidade mais, en xeral, non tal.

REFERENCIAS

- Azorín Poch, F. (1986) *Estadística, información, organización e sistema*. Centro de Información Estadística de Galicia, Santiago de Compostela
- Gómez Duaso, J. (2018) «Tablas multidimensionales en R». *XX Jornadas de Estadística. Comunidades Autónomas*. Logroño
- IGE (2018) Carga de poboación estacional dos concellos de Galicia http://www.ige.eu/web/mostrar_actividade_estadistica.jsp?idioma=gl&codigo=0201001008
- IGE (2017) Manual do paquete Demogal <http://www.ige.eu/estatico/pdfs/s6/demogal-manual.pdf>
- Iglesias Patiño, Carlos L. (2004) «El empleo de los modelos en la estadística oficial». *XIII Jornadas de Estadística. Comunidades Autónomas*. Toledo
- Iglesias, C. e López, E. (2018) «Análise exploratoria da distribución espacial dos centenarios da Galiza». *Actas do III Encontro Luso-Galaico de Biometría*. SPE. Aveiro
- Leguina Herrán, J. (1981) *Fundamentos de demografía*. (3ª edición revisada). Siglo XXI de España Editores. Madrid
- López, E. e Iglesias, C. (2017) «Estratexia de xeorreferenciación da poboación de Galicia empregando técnicas estatísticas» *Actas do XIII Congreso Galego de Estatística e Investigación de Operacións*. SGapeio. Ferrol
- Martín, M et al. (2013) «Las tablas de mortalidad de Galicia y la Región Norte de Portugal empleando el software R». *Actas do I Encontro Portugués de Biometría & I Encontro Luso-Galaico de Biometría*, Braga.
- Teijeiro Alfonsín, E. (1991) «Algunas técnicas multivariantes útiles para la presentación de los resultados de una encuesta». *Estadística Española*, 127, pp 305-324. INE
- Torrado, P. e López, E. (2017) «Carga de poboación estacional dos concellos de Galicia». *Actas do XIII Congreso Galego de Estatística e Investigación de Operacións*. SGapeio. Ferrol

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

O INVENTARIO DE OPERACIÓNS E ACTIVIDADES ESTATÍSTICAS. INDICADORES DE CALIDADE

María Martín Vila¹, Estefanía Villar Cheda¹ e Carlos Iglesias Patiño¹

¹ Instituto Galego de Estatística

RESUMO

O inventario de actividades estatísticas pretende recoller de xeito sistemático, estruturado e homoxéneo as principais características das operacións e actividades estatísticas realizadas pola administración da comunidade autónoma, permitíndonos coñecer a actividade estatística realizada polo sistema estatístico galego. Con este propósito, a base de datos que comprende o inventario recolle, por unha parte, toda a información técnica das operacións e actividades estatísticas (periodicidade, forma de recollida,...) e pola outra, toda a información necesaria para a xestión dos programas estatísticos anuais (data de publicación dos resultados, estado da operación estatística,...) e a súa avaliación. Este estudo revisa a evolución e características das principais variables recollidas no inventario así como as posibles interaccións entre elas co propósito de detectar padróns de comportamento que poderían terse en conta nas programacións ou seguimentos. Entre outras, estúdase a distribución da publicación de resultados no tempo que presenta unha acumulación de difusión nos meses de setembro e decembro, ou a relación entre a tipoloxía da operación ou actividade estatística e a orixe de dos datos, que resulta significativa e forte.

Para o seguimento e a planificación é necesario dispor dunha serie de indicadores que permitan avaliar a produción estatística. Neste contexto e en sintonía co código de boas prácticas europeo elabóranse os indicadores de calidade que entraron a formar parte do inventario no ano 2016. Neste senso, este traballo recolle un estudo do comportamento destes indicadores e relacións entre eles. Entre outros estúdase a relación entre a execución e o calendario ou a posibilidade de facer estimacións cos propios valores pasados do indicador de puntualidade estrutural. Estúdanse tamén os indicadores cualitativos e se teñen relación coa tipoloxía no caso da dispoñibilidade de notas metodolóxicas.

Palabras e frases chave: Estatística pública, Inventario, Indicadores, Calidade

1. INTRODUCCIÓN

O artigo 27.6 do Estatuto de autonomía de Galicia recolle a competencia exclusiva da estatística para os fins da Comunidade Autónoma. En aplicación deste artigo, o Parlamento aprobou a Lei 9/1988, de estatística de Galicia (LEG), modificada polas leis 7/1993, do 24 de maio; 10/2001, do 17 de setembro; 16/2006, do 27 de decembro, e 8/2011, do 9 de decembro. Nela establécese que o Plan galego de estatística (PGE) é o instrumento de ordenación e planificación desta actividade, mediante a progresiva constitución do Sistema estatístico, entendendo por tal o conxunto ordenado e integrado de métodos, procedementos e resultados dos seus axentes institucionais.

Para os efectos da lei de estatística enténdese por actividade estatística galega a recompilación, elaboración e ordenación sistemática da información cuantificable, a publicación e difusión de resultados necesarios ou útiles para o coñecemento cuantitativo, a análise da realidade demográfica, agraria, pesqueira, industrial, comercial, financeira, de servizos, social, cultural e ambiental de Galicia e, en xeral, calquera cuestión referida ás condicións de vida, fins e competencias da comunidade autónoma.

No ano 2016 aprobouse o plan vixente como un sistema de planificación por obxectivos que permite crear, organizar e difundir coñecemento estatístico, acorde coas necesidades das institucións públicas, dos axentes económicos e sociais e mais da cidadanía. Para conseguilo, fórmulanse obxectivos de tres tipos: informativos, instrumentais e de calidade.

O inventario de operacións (OE) e actividades estatísticas (AE) pretende recoller de xeito sistemático, estruturado e homoxéneo as principais características das operacións ou actividades estatísticas que se realizan pola administración da comunidade autónoma, permitíndonos coñecer a actividade estatística realizada polo sistema estatístico galego. O Instituto Galego de Estatística (IGE) elabora os proxectos dos programas estatísticos anuais, artigo 9.2 da Lei de Estatística de Galicia (LEG) e os seus correspondentes informes de seguimento que lle eleva ao Consello Galego de Estatística (CGE), polo que se fai necesario ter información estruturada das operacións e actividades estatísticas incluídas nos citados programas para a súa xestión.

2. INVENTARIO DE OPERACIÓNS E ACTIVIDADES ESTATÍSTICAS

O inventario recolle nunha base de datos, por unha parte, toda a información técnica das OES e AES (periodicidade, forma de recollida,...) e pola outra, toda a información necesaria para a xestión dos programas estatísticos anuais (data de publicación dos resultados, estado da operación estatística,...) e a súa avaliación. A operación ou actividade estatística é a unidade básica do inventario, que abreviaremos como OAE.

As OAE do plan clasifícanse en operación ou actividade estatística segundo a implicación do departamento que a elabora (participa nalgúna fase a maiores da de difusión) e se dispón ou non de proxecto técnico.

O inventario artículase sobre unha base de datos en MySQL a partir da cal extráese a información pertinente tanto para a súa visualización na web do IGE como para a planificación e o seguimento a cargo do servizo de coordinación.

Desde a páxina principal do IGE pódese acceder ao inventario web, cuxo obxectivo é dispoñer dunha aplicación de visualización do inventario de OAE da Comunidade Autónoma coas seguintes funcionalidades:

- Visualización e acceso ás OAE segundo departamentos responsables e segundo os obxectivos do plan (parte identificativa)
- Reflectir as principais características das OAE
- Reflectir a evolución na planificación das OAE tanto ao longo do plan, é dicir nos diferentes programas, como en plans anteriores
- Introducir metainformación das OAE como cuestionarios e proxecto técnico
- Introducir información do seguimento das OEA, tanto dos indicadores clásicos como dos indicadores de calidade da planificación

O acceso ao inventario pódese realizar ben por organismo responsable ou ben por obxectivos informativos (segundo orde: metas de información, perspectiva de xénero ou liñas de actuación salientables) o que permite que o inventario non sexa unha simple listaxe de OAE senón que pretende ser unha ferramenta de consulta estruturada e informativa.

A parte identificativa de cada OAE consta do organismo ou organismos responsables e fases que elaboran, a tipoloxía e estado, os obxectivos e os obxectivos informativos. Respecto da tipoloxía cabe resaltar que no último plan de estatística homoxeneizáronse as operacións e actividades e elimináronse as operacións en proxecto polo que as posibilidades para a tipoloxía son OE ou AE e o estado, no caso das operacións pode ser: en implantación, en reestruturación, en curso ou en cumprimento. Ao comezar o PGE2017-2021, no ano 2017, o inventario constaba de 148 OAE e incrementouse en 11 OAE ata chegar ás 159 no inventario actual (2019). Por tipoloxía e estado temos a seguinte distribución:

Tipo	Estado	Consolidadas	Inventario 2017	Inventario 2019
OE	en reestruturación	1	2	1
OE	en implantación	0	3	3
OE	en curso	65	69	72
OE	en cumprimento	10	13	13
Total OE		76	87	89
AE		41	61	70
Total		117	148	159

Táboa 1: Número de OAE segundo tipoloxía (OE/AE) e estado.

A evolución do inventario segue o padrón que se espera, é dicir, hai un incremento nas OE en curso, froito de que tres das 5 operacións novas no 2017 (en reestruturación ou en implantación) evolucionaron a en curso. E o incremento maior (6%) experimentouse nas actividades estatísticas, que pasaron de 61 no 2017 a 70 no inventario actual.

No apartado de características técnicas móstranse as principais características propias de cada OAE, como son a forma de recollida da información, as variables de estudo, unidades informantes... que non varían co ano de programación e móstranse referidas sempre ao estado actual da OAE.

O 20% das OAE inventariadas son do tipo mostraxe, exhaustiva ou toma directa de datos, outro 20% de síntese e o maior peso recae nas estatísticas que se refiren a explotacións, o 60%, xa sexan de microdatos ou de sistemas de información.

Ademais das metas, o PGE determina unha listaxe de operacións e actividades “consolidadas” establecendo pois un conxunto mínimo de estatísticas básicas e polo tanto relevantes, se temos en conta só as OAE consolidadas o padrón de distribución consérvase e a evolución do inventario no que levamos de plan non reflicte anomalías xa que se incrementaron nun 4% as estatísticas de explotación e nun 3% as de síntese (en concreto as de contas económicas e balances), isto está en liña co artigo 5 do PGE que exige ás estatísticas a adecuación aos criterios de economía e de aproveitamento das fontes existentes e a minimización das molestias ás e aos informantes. As demais tipoloxías mantivéronse estables ou non tiveron incrementos significativos.

O seguinte apartado das características técnicas son os Ámbitos de estudo. Dentro deles detállanse o alcance, as variables de estudo, as variables de clasificación (que se mostran se están definidas) e a periodicidade. Na análise do inventario segundo a periodicidade da OAE é preciso ter en conta que pode ter máis dunha periodicidade, é dicir, unha mesma OAE poder ter periodicidade anual e trimestral se difunde datos trimestrais pero á vez fai unha revisión anual (considérase para a análise a máis curta). O 65% das OAE teñen periodicidade regular anual, o 8% son conxunturais (periodicidade menor que semestral) e o 8% son irregulares. A evolución, respecto ás OAE inventariadas no 2017, ao inicio de PGE2017-2021, non mostra anomalías posto que medraron as OAE con periodicidade mensual e anual e as irregulares. Isto último ten explicación porque os módulos da Enquisa Estrutural a Fogares e da Enquisa de Poboación Activa teñen periodicidade irregular, así mesmo, recoméndase para as actividades estatísticas a programación con este periodicidade (ou esporádica) cando son actividades novas e propónse modificala unha vez que a actividade ten percorrido suficiente para asegurar a súa periodicidade na difusión.

Cómpre neste intre facer unha diferenza entre a periodicidade na difusión e a dos datos ofrecidos pois non sempre coinciden. O calendario de difusión recóllese noutra variable do inventario e publícase no decreto de programa pero non se mostra no inventario web, ademais esta variable só se rexistra para as incluídas en cada programa. Así, das 144 OAE programadas no ano 2019, 10 teñen actualización continua ou en varias datas non establecidas de antemán polo que non se lle pode atribuír un mes preciso de difusión.

O estudo das datas de difusión das estatísticas periódicas (non teremos en conta as irregulares, de actualización continua ou en varias datas nin as regulares cada cinco anos) mostra unha clara acumulación no mes de setembro das estatísticas de periodicidade regular anual. Tamén no mes de decembro hai certa acumulación de publicación destas estatísticas debido a que cando non se ten unha data establecida prográmase a difusión para final de ano co fin de non penalizar o calendario e refinalo co tempo. Os meses de xaneiro e agosto son claramente os meses de menor difusión debido ao comezo do ano e o período de vacacións, isto fai que febreiro, e como xa dixemos setembro, teñan maior carga. As estatísticas de periodicidade mensual son estables, como cabería esperar e as trimestrais están máis concentradas nos meses 1,4,7 e 10 nos que se publican 7 das 12 estatísticas de deste tipo.

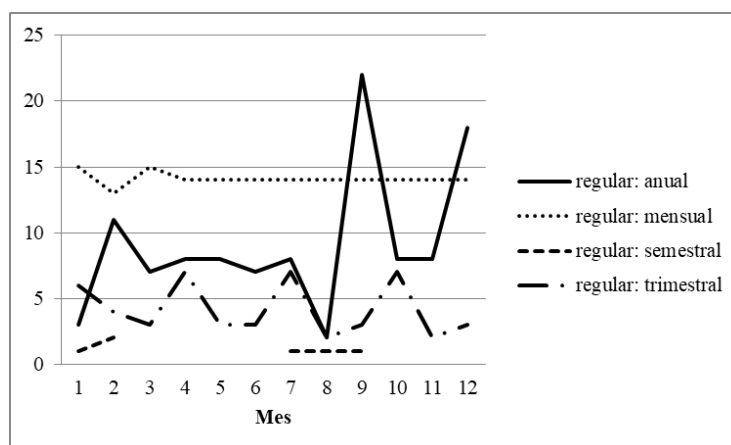


Figura 1: Calendario de difusión das OAE regulares.

Na procedencia da información rexístranse a orixe dos datos, a unidade estatística elemental, as outras unidades estatísticas, a unidade informante, as formas de cobertura, recollida e obtención da información (facilitada pola fonte, captura da información publicada...) e a variable proporcionada pola fonte. O tipo de información mostrada dependerá da tipoloxía da OAE e a orixe dos datos. Esta información permite coñecer a trazabilidade dos datos.

No estudo da tipoloxía da OAE segundo a orixe dos datos, agrupando as categorías dunha forma razoada temos:

	Cuestionarios/ medicións	Microdatos/ rexistros e sistemas de información administrativos	Formularios, resumos e estados/información elaborada	Diferentes tipos de orixe	Total
Mostraxe/Exhaustiva/ toma directa de datos	32	0	0	1	33
Explotación	1	55	26	11	93
Síntese	1	3	8	21	33
Total	34	58	34	33	159

Táboa 2: Número de OAE segundo tipoloxía e orixe dos datos.

Sobre esta táboa de continxencia calculamos o estatístico khi-cadrado para ver se a relación de dependencia que se mostra coa simple visualización é real. Para medir a asociación entre as variables empregamos a función `assocstats()` do paquete R vcd:

	X ²	df	P(> X ²)
Likelihood Ratio	178.46	6	0
Pearson	190.09	6	0

Phi-Coefficient : NA

Contingency Coeff.: 0.738

Cramer's V : 0.773

que ademais de ofrecer o test khi-cadrado, claramente significativo para rexeitar a independencia, proporciona o valor do estatístico V de Cramer. Segundo este estatístico a relación entre as dúas variables é moi forte (Cramer's V > 0,5).

Se analizamos o inventario segundo o máximo nivel de desagregación que ofrece a OAE temos que o 19% das estatísticas inventariadas ofrecen datos a nivel de concello, preto do 30% desagregan a información por provincias e máis do 30% ofrecen datos de Galicia como máxima desagregación. Así mesmo, o 12 % das estatísticas inventariadas ofrecen datos con outra desagregación, xa sexa sección ou distrito censual ou unha clasificación específica.

Se estudamos a relación entre o nivel de desagregación (en grupos agregados) e a tipoloxía da OAE temos a seguinte táboa de continxencia:

	Galicia	Menor que Galicia e maior que concello	Concello	Menor que concello
Enquisas, exhaustivas e toma directa	12	13	3	5
Explotacións	20	37	23	13
Síntese	21	8	4	0

Táboa 3: Número de OAE segundo a tipoloxía e o nivel de desagregación.

	X ²	df	P(> X ²)
Likelihood Ratio	26.524	6	0.00017777
Pearson	23.701	6	0.00059257

Phi-Coefficient : NA
 Contingency Coeff.: 0.36
 Cramer's V : 0.273

O test proporciona unha khi-cadrado significativa e a V de Cramer mostra un valor medio de asociación entre ambas variables. Cabe mencionar que o maior porcentaxe de estatísticas con desagregación de “concello e menor” provén de explotacións de datos así como que ningunha estatística de síntese proporciona datos cunha desagregación menor que concello, o que era esperable polas características propias das estatísticas de síntese.

No apartado de planificación estatística recóllese a información da OAE en cuestión nos anteriores plans e no actual. Así, unha parte é histórica (plans anteriores) e outra vai ir cambiando cos programas do plan actual, engadíndose cada ano un novo ítem. Neste apartado móstranse as OEA tal e como foron programadas polo que permite seguir tanto a evolución respecto á súa tipoloxía e o estado, no caso de ser unha operación, como a dependencia orgánica.

3. INDICADORES DE CALIDADE

O código de boas prácticas das estatísticas europeas constitúe un marco común de calidade para o Sistema Estatístico Europeo. Nel enuméranse un total de 16 principios, que abarcan o entorno institucional, os procesos de elaboración de estatísticas e a produción estatística.

Neste contexto, procúranse calcular os indicadores relacionados cos 5 principios relativos á produción estatística: pertinencia, precisión e fiabilidade, oportunidade e puntualidade, coherencia e comparabilidade e accesibilidade e claridade. Comezouse calculando os indicadores que son factibles segundo o establecido na ESS Guidelines for the implementation of the ESS Quality and performance indicators (QPI) e continuarase engadindo outros axustados á realidade do sistema estatístico galego. Deste xeito, os indicadores noméanse cunha ou dúas letras relacionadas co principio ao que fan referencia (na súa denominación inglesa) seguida/s de números correlativos. No PGE 2017-2021 introdúcense tamén estes cinco principios como parte dos obxectivos de calidade das estatísticas, polo que estes indicadores de calidade empregáranse para avaliar o devandito plan. Ademais inclúense os indicadores clásicos que se veñen empregando para o seguimento dos programas: Execución e calendario.

A execución dunha operación mídese comprobando os resultados acadados en cada fase, desde a recollida de datos ata a difusión. A execución das actividades estatísticas calcúlase en función dos resultados das actuacións concretas que figuran no anexo III do decreto de programa, e de xeito similar ás operacións, valóranse as distintas fases de forma implícita.

Para cada organismo e para o global da comunidade autónoma o grao de execución das operacións obtense como media ponderada das operacións correntes (80 %) e as novas (20%), e o grao de execución global obtense como media ponderada do grao de execución das operacións -OE- (60 %) e das actividades -AE- (40 %). No último programa avaliado, PEA2018, a execución global foi do 97,6%. As OE correntes acadaron unha execución do 96,7% e un 100% as novas mentres que as actividades acadaron unha execución do 98%. Se analizamos máis polo miúdo, o 92% das OAE acadaron o máximo nivel de execución (100%), o 3% non chegaron ao 100 pero superaron o 90% e só dous OAE (unha operación en curso e unha actividade) tiveron unha execución do 0%.

O indicador de calendario mide a puntualidade na publicación de resultados. Refírese ás operacións e actividades con difusión prevista no programa, sen considerar as de tipo indicadores que se actualizan continuamente segundo o permiten as fontes primarias. En cada OAE compróbase se se publicaron os datos no soporte e data previstos e calcúlase un indicador de calendario global como media ponderada do indicador de puntualidade das operacións -OE- (60%) e das actividades -AE- (40 %). No último programa avaliado, PEA 2018, o indicador global de calendario estivo por riba do 77%, as operacións acadaron preto do 70% mentres que as actividades case chegan ao 87%.

A evolución na execución das OE correntes mostra que nos atopamos no máximo nivel de execución desde o ano 2007 e repunta despois duns anos de baixa execución en termo globais. Pola súa parte a evolución do indicador de calendario das operacións correntes mostra un lixeiro ascenso con respecto aos últimos programas pero aínda non se recupera o máximo acadado no ano 2013.

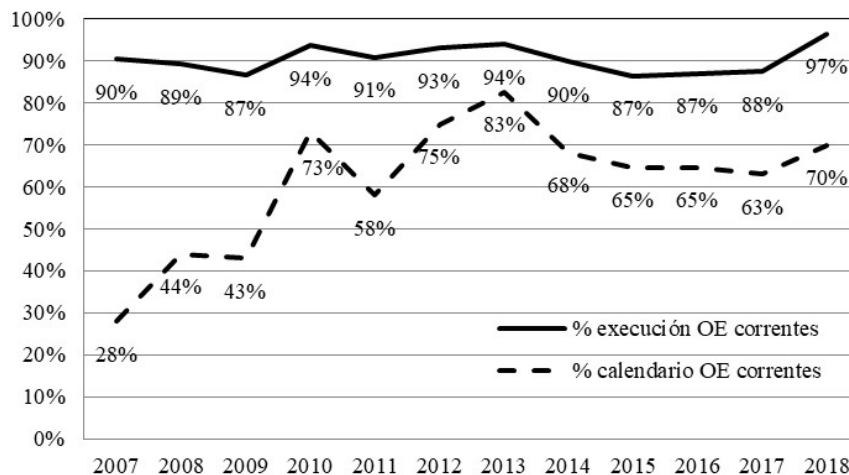


Figura 2: Evolución da execución e o calendario das OE correntes.

Podería pensarse que estas dúas variables están relacionadas e incluso co número de OE correntes programadas. Baixo este suposto estimamos tres modelos de regresión para o calendario en función da execución, considerando todo o período dispoñible e os datos dos últimos plans (2012 en diante) e o calendario en función da execución e o número de operacións correntes programadas. Ningún destes modelos resultou significativo, e o que parece ao principio sorprendente é un bo síntoma de cara a avaliación da actividade estatística posto que canto máis independentes sexan os indicadores máis aspectos da estatística se avalían.

Os indicadores de oportunidade tratan de amosar o tempo transcorrido entre o período de referencia do dato e a data de publicación dun valor referido ao devandito período. O indicador TP1 calcúlase coa primeira data de publicación dos datos provisionais referidos a un período; pola contra, o indicador TP2 calcúlase coa data de publicación dos datos definitivos dese período. O valor numérico destes indicadores será o número de meses transcorridos entre a referencia do dato e da data de publicación (aínda que esta sexa anterior á establecida na programación anual). Xa que estes indicadores se calculan como o número de meses transcorridos, separáranse as operacións estruturais (aquelas cuxa periodicidade sexa igual ou superior á semestral) das operacións conxunturais (aquelas cuxa periodicidade sexa igual ou inferior á cuatrimestral). Cando hai varias datas de publicación a avaliar no PEA, o indicador final será o resultado da media de cada indicador avaliado en cada publicación. Procederáse tamén a calcular o valor medio cando unha operación ou actividade conteña varias series principais nas que se avalía este indicador, considerando este promedio como o valor do indicador asignado á operación ou actividade. Nos casos nos que na data de avaliación do programa estatístico anual haxa algún/s dato/s pendente/s de publicación, a efectos do cálculo destes indicadores, supónse que se publicou/publicaron na devandita data de avaliación. De ocorrer isto, indicárase que o indicador calculado deste xeito é unha estimación.

O indicador de oportunidade global obtense como media ponderada do indicador de puntualidade das operacións -OE- (60%) e das actividades -AE- (40 %). No último programa avaliado este indicador foi de 13 meses para as estatísticas estruturais e 3 meses para as conxunturais. Desde que comezou a calcularse este indicador separadamente para estruturais e conxunturais, no seguimento do PEA2015, a oportunidade das OE correntes mantense estable.

Resulta interesante a estimación do valor do TP2 estrutural dun ano a partir do ano precedente, para estudar esta relación propónse dous modelos, un primeiro que inclúe unicamente os valores que non se estiman en ambos os dous anos que consideramos, os dous últimos dispoñibles (68 OAE)

```
lm(formula = Valor18 ~ Valor17, data = DatosNE)
Residuals:
    Min     1Q   Median     3Q      Max
-6.1801 -0.5595  0.1045  0.6025  9.5354
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84406   0.57033   1.48   0.144
Valor17      0.90515   0.05046  17.94 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.586 on 66 degrees of freedom
Multiple R-squared:  0.8298,    Adjusted R-squared:  0.8272
F-statistic: 321.8 on 1 and 66 DF, p-value: < 2.2e-16
```

E outro que ten en conta todos os valores (reais e estimados) do TP2 estrutural (90 OAE)

```
lm(formula = Valor18 ~ Valor17, data = Datos)
Residuals:
    Min     1Q   Median     3Q      Max
-6.9639 -0.7496 -0.4128  1.3117  9.5872
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29030   0.59671   0.486   0.628
Valor17      1.03062   0.04434  23.242 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.944 on 88 degrees of freedom
Multiple R-squared:  0.8599,    Adjusted R-squared:  0.8583
F-statistic: 540.2 on 1 and 88 DF, p-value: < 2.2e-16
```

Os dous modelos propostos resultan significativos e cunhas correlacións moi altas polo que cabería reformular o xeito de estimar o TP2 no caso de non terse publicada a OAE.

Se consideramos o seu homólogo conxuntural a relación entre os datos dos dous anos é “perfecta” se non temos en conta os datos estimados (unha OAE) e cuasi perfecta se temos en conta todas as OAE para as que se calcula. Neste caso hai que ter en conta que as observacións son menos xa que este indicador se calcula para menos OAE (19).

O indicador de puntualidade TP3 indica o tempo de retraso transcorrido entre a data de publicación prevista duns resultados estatísticos e a data de publicación efectiva deses resultados. O valor numérico destes indicadores será o número de meses entre a data de publicación prevista e da data de publicación real, e polo tanto será un número enteiro. Ao igual que nos indicadores TP1 e TP2 nos casos nos que na data de avaliación do programa estatístico anual haxa algún/s dato/s pendente/s de publicación estimárase o valor e supónse que se publicou/publicaron na devandita data de avaliación. Así mesmo, cando hai varias datas de publicación a avaliar no PEA, o indicador final será o resultado da suma de cada indicador avaliado en cada publicación. No último programa avaliado, PEA2018, o 72% das 132 OAE para as que se calcula este indicador obtiveron o mínimo retraso, 0 meses. Das OAE que difundiron datos, indicador non estimado, o retraso medio foi de 4 meses, mentres que as OAE que non difundiron datos acumulan un retraso medio de 40 meses (estimado).

O indicador TP4 reflectirá a porcentaxe de difusión incluídas na programación anual que, efectivamente, se axustaron ao calendario deseñado e indica a taxa de puntualidade das publicacións previstas para o programa estatístico anual que se analiza. O valor numérico destes indicadores será o cociente entre o número de publicacións que se difundiron segundo o calendario establecido e o total de publicacións previstas na programación anual e, posteriormente, multiplicado por 100.

A diferenza entre este indicador de puntualidade e o indicador de calendario radica en que no indicador de calendario outórgase puntuación ás estatísticas anuais que publican con un ou dous meses de retraso mentres que, se a OAE só ten esta publicación e non publican no mes establecido a % de puntualidade será 0. Por outra banda, neste indicador tense en conta as publicacións pendentes de

programas anteriores coma en todos os TP mentres que no calendario só se ten en conta o publicado no decreto de programa do ano avaliado. Se analizamos a relación entre ambos indicadores atopamos unha correlación moi alta, que non sorprende:

```
lm(formula = TP4 ~ IGE_Calendario, data = segCalTP4)
Residuals:
    Min     1Q   Median     3Q      Max
-71.622   2.828   2.828   2.828   5.029
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.02947   1.24569  -4.037 6.53e-05 ***
IGE_Calendario  1.02202   0.01474 69.358 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 12.2 on 381 degrees of freedom
Multiple R-squared:  0.9266,    Adjusted R-squared:  0.9264
F-statistic: 4810 on 1 and 381 DF, p-value: < 2.2e-16
```

Ademais deste modelo, no que se inclúen todos os datos dispoñibles (período 2015-2018) para todas as OAE avaliadas, formulamos outros modelos para avaliar o impacto do ano de programación na relación. Así detectamos que o ano 2015 é anómalo posto que foi o primeiro ano no que se avaliou o indicador TP4 e os cálculos fixéronse a posteriori da avaliación ordinaria o que puido xerar interferencias:

```
lm(formula = TP4 ~ D16 + D17 + D18 + IGE_Calendario + I(D16 *
  IGE_Calendario) + I(D17 * IGE_Calendario) + I(D18 * IGE_Calendario),
  data = segCalTP4)
Residuals:
    Min     1Q   Median     3Q      Max
-71.434   1.122   1.895   3.292  19.205
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -19.20530   3.19747  -6.006 4.49e-09 ***
D16           17.01502   4.09781   4.152 4.08e-05 ***
D17           18.08358   3.81569   4.739 3.05e-06 ***
D18           14.82006   3.89685   3.803 0.000167 ***
IGE_Calendario    1.14480   0.03953  28.959 < 2e-16 ***
I(D16 * IGE_Calendario) -0.13382   0.05009  -2.672 0.007876 **
I(D17 * IGE_Calendario) -0.15253   0.04655  -3.277 0.001149 **
I(D18 * IGE_Calendario) -0.13387   0.04713  -2.840 0.004754 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.85 on 375 degrees of freedom
Multiple R-squared:  0.9319,    Adjusted R-squared:  0.9307
F-statistic: 733.4 on 7 and 375 DF, p-value: < 2.2e-16
```

Se eliminamos do cálculo os datos do ano 2015 temos unha relación moi boa, no senso dun R-cadrado moi elevado (que xa tiñamos) pero coa diferenza de que o efecto do ano de programación deixa de ser significativo, tanto aditiva como multiplicativamente.

```
lm(formula = TP4 ~ D17 + D18 + IGE_Calendario + I(D17 * IGE_Calendario) +
  I(D18 * IGE_Calendario), data = segCalTP4_)
Residuals:
    Min     1Q   Median     3Q      Max
-71.434   1.122   1.895   3.292   4.385
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.190e+00  2.260e+00  -0.969   0.333
D17           1.069e+00  2.912e+00   0.367   0.714
D18          -2.195e+00  2.995e+00  -0.733   0.464
IGE_Calendario    1.011e+00  2.713e-02  37.265 < 2e-16 ***
I(D17 * IGE_Calendario) -1.871e-02  3.473e-02  -0.539   0.590
I(D18 * IGE_Calendario) -4.899e-05  3.533e-02  -0.001   0.999
```

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 10.45 on 315 degrees of freedom
 Multiple R-squared: 0.9457, Adjusted R-squared: 0.9448
 F-statistic: 1097 on 5 and 315 DF, p-value: < 2.2e-16

Co indicador R1 procurarase medir a relevancia dunha actividade en canto ao seu interese e, dado que o PGE é o marco de actuación do sistema estatístico galego, procuraremos cuantificala establecendo uns criterios en termos deste plan. Así pois, reflectiremos o grao de interese dunha OAE a partir dunha variable categórica que recolla maiores valores cando a OAE:

- satisfaga unha meta ou perspectiva que non se cumpre con outras operacións ou actividades,
- sexa consolidada, e/ou
- achegue información estatística sobre Galicia non dispoñible noutras fontes.

e, asignaráselle ao indicador R1 o valor da suma dos puntos obtidos nos tres criterios de relevancia analizados para a OAE e que será polo tanto máis elevado canta máis relevancia acade a OAE.

O indicador do grao de relevancia implantouse na avaliación do último programa do plan 2012-2016 polo que temos dispoñible neste intre a serie 2016-2018. Por outra banda, para o ano 2016 só se consideran as OAE que permanecen no inventario no seguinte plan:

R1	2016	2017	2018	F16	F17	F18
0	0	1*	0	0,000	0,007	0,000
1	67	28	33	0,486	0,212	0,229
2	52	74	75	0,862	0,752	0,750
3	17	29	32	0,986	0,964	0,972
4	1	3	2	0,993	0,985	0,986
5	1	2	2	1,000	1,000	1,000

Táboa 3: Número de OAE e frecuencia acumulada segundo a relevancia e o ano de programación.

O cambio de plan do ano 2016 ao 2017, e polo tanto o novo proceso de consolidación, ten un efecto directo sobre este indicador posto que é unha das compoñentes. A evolución lóxica e esperada do indicador é que as operacións se vaian acumulando nas categorías 1 e 2 posto que ao ir engadindo operacións novas o valor asociado ao primeiro criterio (número de metas que unha operación ou actividade satisfai en exclusiva) pode ir diminuíndo, mentres que os outros dous criterios son fixos no plan a non ser que unha OAE deixe de difundir a información nova que difunde, pero este é un requisito para programarse como OAE polo que tería que estudarse detidamente a súa programación.

Polo de agora só dispomos dun ano fin de plan, 2016. O ano 2018 domina estocasticamente ao 2016, abonda ollar para as columnas de frecuencias relativas acumuladas, $F18 < F16$. O 2017 case domina tamén ao 2016, agás na anomalía do cero.

Por último, dentro dos indicadores de accesibilidade e claridade (AC), temos os indicadores AC1e AC2. O AC1 amosa a existencia ou non dun punto de información (páxina web, publicación impresa, pendrive, etc.) onde se recolla toda a información da operación ou actividade estatística en cuestión, pero exclusivamente esa información. É dicir, perséguese unha total correspondencia entre o punto de difusión e a operación ou actividade estatística. Os valores que pode tomar este indicador son dous: "ten"/ "non ten" e das 144 OAE programadas no 2018 o 80% ten ligazón única e exclusiva.

O indicador AC2 especifica se, no espazo de difusión da estatística ou na propia publicación, se xuntan unhas notas metodolóxicas que expliquen os conceptos principais da estatística (ámbitos, definicións...), as fontes empregadas para a elaboración da estatística, o proceso de cálculo e o conxunto de resultados que se van difundir. Igual que o indicador anterior este indicador é unha variable dicotómica que só ten dous posibles resultados: "posúe"/"non posúe". Para o último programa avaliado, 2018, o 58% das OAE posúen notas metodolóxicas. Cabe remarcar que 82 das 144 son operacións estatísticas en curso (con proxecto técnico) polo que bastaría con enlazar os resultados co proxecto no mesmo sitio de difusión para acadar un bo resultado no indicador. Todas as operacións responsabilidade do IGE, 26 das 84

* Operación en implantación que non chegou a implantarse (enquisa a empresas multilocalizadas)

operacións, por proceder do Instituto e pola propia arquitectura da páxina web teñen a ligazón ao proxecto técnico accesible e polo tanto posúen notas metodolóxicas.

Notas metodolóxicas	OE	AE	Total
Non posúe	27	34	61
Posúe	57	26	83

Táboa 4: Número de OAE segundo o valor do AC2 e a tipoloxía da OAE.

Para analizar a relación entre a tipoloxía das OAE (operación ou actividade) e o valor do indicador AC2 empregamos o test exacto de Fisher, que resulta significativo, polo que non podemos asegurar independencia entre as variables estudadas:

```
Fisher's Exact Test for Count Data
data: tta
p-value = 0.003849
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1722374 0.7599686
sample estimates:
odds ratio
0.3648967
```

E se analizamos o nivel de asociación co coeficiente V de Cramer temos que a relación non é demasiado intensa (Cramer's $V < 0,3$)

	X ²	df	P(> X ²)
Likelihood Ratio	8.6501	1	0.0032705
Pearson	8.6211	1	0.0033230

```
Phi-Coefficient : 0.245
Contingency Coeff.: 0.238
Cramer's V      : 0.245
```

REFERENCIAS

Código de buenas prácticas de las estadísticas Europeas. Eurostat 2017. ISBN 978-92-79-80025-2 URL <https://ec.europa.eu/eurostat/documents/4031688/9394048/KS-02-18-142-ES-N.pdf/e792b761-6f09-42a9-a1e0-3a3356a0de1c>

ESS handbook for quality reports. Eurostat 2015. ISBN 978-92-79-45487-5 URL <https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf>

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Inventario de operacións e actividades estatísticas. DEFINICIÓNS. Instituto Galego de Estatística URL: http://www.ige.eu/estatico/pdfs/s5/inventario_actividades_estadisticas/definicions_inventarioN.pdf

Indicadores de calidade das operacións estatísticas. METODOLOXÍA. Instituto Galego de Estatística URL: http://www.ige.eu/estatico/pdfs/s5/inventario_actividades_estadisticas/MetodologiaIndicadoresCalidade.pdf

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

PROXECCIÓNS DE TAXAS DE ACTIVIDADE

Torrado Solar, Pablo¹ e Alberte Aparicio, Eva¹

¹ Instituto Galego de Estatística

RESUMO

As implicacións macroeconómicas do envellecemento que se espera que experimente a poboación nas próximas décadas dependerán en boa medida dos recursos de forza de traballo dos que dispoña a economía no futuro. Así pois, os encargados de deseñar as políticas públicas teñen un interese cada vez maior por dispoñer dunha simulación estatística do comportamento futuro da poboación en relación co mercado laboral. O obxectivo deste traballo é expoñer a metodoloxía que emprega o Instituto Galego de Estatística para proxectar as taxas de actividade xunto co volume e estrutura da poboación activa por sexo e idade.

Palabras e frases chave: taxas de actividade, ratios de actividade, modelos de extrapolación loxística, transformacións loxísticas, tendencias temporais

1. INTRODUCCIÓN

A *Enquisa de poboación activa (EPA)* define a taxa de actividade como a porcentaxe de activos respecto á poboación de 16 ou máis anos, sendo a poboación activa o conxunto de persoas de 16 ou máis anos que subministran man de obra para a produción de bens e servizos ou que están sen traballo, dispoñibles para traballar e buscando activamente un emprego. Así pois, a poboación activa representa os recursos de forza de traballo dos que dispón unha economía e a taxa de actividade mide a propensión das persoas en idade de traballar a participar no mercado laboral.

A proxección do comportamento futuro da poboación en relación co mercado laboral resulta fundamental na planificación das políticas públicas, xa que é determinante no crecemento potencial da economía, sustentabilidade dos sistemas públicos...

Se para o ano t , sexo s ($s=homes, mulleres$) e grupo quinquenal de idade i ($i=16-19, 20-24, \dots, 65-69, 70+$) denotamos:

$P_{s,i}^t$: poboación no ano t do sexo s e grupo quinquenal de idade i

$TA_{s,i}^t$: taxa de actividade no ano t da poboación de sexo s e grupo quinquenal de idade i

temos que para calquera grupo poboacional G que sexa agregación de grupos elementais, a taxa de actividade e a poboación activa no ano t veñen dadas por:

$$TA_G^t = \sum_{(s,i) \in G} TA_{s,i}^t \cdot \frac{P_{s,i}^t}{P_G^t}$$

$$PA_G^t = TA_G^t \cdot P_G^t = \sum_{(s,i) \in G} TA_{s,i}^t \cdot P_{s,i}^t$$

onde P_G^t denota a poboación no ano t do grupo poboacional G :

$$P_G^t = \sum_{(s,i) \in G} P_{s,i}^t$$

Polo tanto, a proxección da taxa de actividade e a poboación activa de calquera grupo poboacional que se obteña como agregación de grupos elementais redúcese a dispoñer das proxeccións de poboación e a proxectar as taxas de actividade correspondentes aos grupos elementais.

As *Proxeccións de poboación a curto prazo* permítenos coñecer a evolución futura do volume e estrutura da poboación por sexo e grupo quinquenal de idade. Agora ben, a poboación obxecto de estudo

nasas proxeccións de poboación é a poboación residente en Galicia a 1 de xaneiro, tanto en vivendas familiares como en establecementos colectivos, pero a EPA vai dirixida á poboación residente en vivendas familiares e obtén os resultados anuais como a media dos catro trimestres. Polo tanto, teremos que adaptar estas proxeccións de poboación ao noso contexto para que fagan referencia á media anual de poboación residente en vivendas familiares.

A proxección das taxas de actividade por sexo e grupo quinquenal de idade farémola a partir das series históricas de taxas de actividade proporcionadas pola EPA. Entre os organismos nacionais e internacionais que elaboran proxeccións de taxas de actividade, existen catro enfoques diferentes para proxectar as taxas de actividade:

- Métodos de xuízo. Un grupo de expertos emite un xuízo sobre o futuro das taxas de actividade en base ás series históricas, datos observados en países próximos cultural e estruturalmente pero máis avanzados economicamente...
- Modelos de regresión. Axústase un modelo de regresión para as taxas de actividade en función de variables explicativas de carácter demográfico, socioeconómico e institucional. Unha vez fixado o escenario futuro para as variables explicativas, utilízase o modelo de regresión axustado para proxectar as taxas de actividade.
- Modelos de extrapolación no tempo. Axústase en función do tempo unha curva sigmoide para as taxas de actividade, normalmente a curva loxística, o que permite extrapolar os seus valores futuros. As curvas sigmoides son moi útiles para modelar procesos nos que se espera alcanzar un estado estable a longo prazo (taxas de actividade, inflación, produtividade...).
- Modelos baseados nunha visión xeracional. A proxección das taxas de actividade realízase a partir das probabilidades de entrada ou saída do mercado laboral correspondentes aos últimos anos, que se estiman considerando a información transversal como lonxitudinal (xeración ficticia).

As oficinas de estatística decántanse sobre todo por modelos de extrapolación no tempo ou modelos baseados nunha visión xeracional, xa que tanto os métodos de xuízo como os modelos de regresión requiren establecer unha opinión sobre o comportamento futuro das taxas de actividade ou das variables explicativas dos modelos de regresión, o cal resulta complexo e subxectivo.

No mercado laboral galego estase producindo un proceso de converxencia das taxas de actividade femininas coas masculinas, polo que resulta especialmente axeitado o uso de modelos de extrapolación no tempo para proxectar as taxas de actividade, pois permiten extrapolar directamente as taxas de actividade masculinas e extrapolar indirectamente as taxas de actividade femininas extrapolando as ratios entre as taxas de actividade femininas e masculinas.

Ao igual que acontece coas proxeccións de poboación, é importante ter en conta que as proxeccións de taxas de actividade constitúen unha simulación estatística da súa evolución futura no caso de manterse as tendencias observadas no pasado. Así pois, non pretenden adiviñar o futuro, senón proporcionar unha ferramenta de apoio para a toma de decisións en base a unha simulación estatística do rumbo que tomaría a forza de traballo baixo a hipótese de que se mantivesen as tendencias pasadas.

2. FONTES DE INFORMACIÓN

As taxas de actividade nos grupos elementais para o período de observación obtémolas da EPA e para calcular a poboación residente en vivendas familiares dos grupos elementais no período de proxección utilizamos as *Proxeccións de poboación a curto prazo*.

2.1. ENQUISA DE POBOACIÓN ACTIVA

A EPA é unha investigación por mostraxe de periodicidade trimestral dirixida á poboación residente en vivendas familiares coa finalidade de coñecer as características desa poboación en relación co mercado laboral.

Desde os seus inicios en 1964, sufriu numerosos cambios na súa metodoloxía e elaboración, entre os que cabe destacar:

- En 1987 introduciuse un novo cuestionario para adaptarse as recomendacións da *Conferencia internacional de estatísticos do traballo* da Organización Internacional do Traballo de 1982, incorporar as variables de estudo da *Enquisa de forza de traballo* da Unión Europea e incorporar información sobre as novidades do mercado laboral español (incremento do traballo a tempo parcial, aumento da contratación temporal...).
- No ano 2002 pasouse a utilizar a definición de parado establecida no *Regulamento 1987/2000 da Comisión das Comunidades Europeas*, do 7 de setembro de 2000.

A modelaxe temporal das taxas de actividade anuais por sexo e grupo quinquenal de idade require dispoñer de series homoxéneas de taxas de actividade anuais, polo que restrinxíndonos aos resultados trimestrais que proporciona a EPA desde 1987 temos que resolver o problema que supón o cambio na definición de parado que se empezou a aplicar en 2002 e converter a información trimestral en información anual.

Segundo a definición de parado que se viña utilizando ata o cuarto trimestre de 2001 clasificábanse como parados a persoas que segundo a nova definición pasan a ser inactivos, polo que para dispoñer de series homoxéneas desde o primeiro trimestre de 1987 hai que corrixir o número de parados e en consecuencia de activos e inactivos anteriores ao primeiro trimestre de 2002.

Para levar a cabo este enlace nas series trimestrais de parados utilizamos o documento *Enlace de las series de paro 1976-2000 según la definición EPA-2002 (J. Trejo e L. Ortega. INE)*, no que estimaron as series trimestrais de parados baixo a nova definición mediante un modelo de regresión probit con variable resposta binaria que indica se unha persoa é parada segundo ambas definicións ou se coa nova definición deixa de ser parada e variables explicativas o sexo, a idade e a interacción entre o sexo e a idade. Para axustar este modelo de regresión probit utilizaron os microdatos da EPA correspondente ao período 2001-2004 no que se recolleu a información necesaria para saber se unha persoa é parada tanto coa antiga como coa nova definición. Esta é a razón pola que realmente só necesitaron recalcular as series trimestrais de parados segundo a nova definición ata o cuarto trimestre de 2000.

Unha vez levado a cabo este enlace xa dispoñemos de series trimestrais homoxéneas de ocupados, parados, activos, inactivos e poboación por sexo e grupo quinquenal de idade desde o primeiro trimestre de 1987. Así pois, podemos obter as series homoxéneas de taxas de actividade anuais desde 1987 dividindo o número medio anual de activos entre a poboación media anual, entendendo que as medias anuais de activos e poboación se calculan como a media aritmética dos catro trimestres.

2.2. PROYECCIONES DE POBOACIÓN A CURTO PRAZO

As proxeccións de poboación por sexo e grupo quinquenal de idade que ofrecen as *Proyecciones de poboación a curto prazo* refírense á poboación residente en Galicia a 1 de xaneiro, tanto en vivendas familiares como en establecementos colectivos.

Agora ben, nos necesitamos proxeccións de poboación por sexo e grupo quinquenal de idade que aludan a media anual de poboación residente en vivendas familiares e que sexan coherentes coas estimacións anuais de poboación que proporciona a EPA para o último ano.

Para obter as proxeccións de poboación que imos utilizar partimos das estimacións anuais de homes e mulleres entre os 16 e 64 anos e 16 ou máis anos que proporciona a EPA para o último ano e proxectamos a media anual de poboación residente en vivendas familiares para eses grupos poboacionais utilizando as taxas de variación que presentan nas *Proyecciones de poboación a curto prazo* despois de transformar a información a 1 de xaneiro en información media anual. Unha vez temos proxectadas as medias anuais de poboación residente en vivendas familiares para os grupos poboacionais considerados, desagregámolas por grupo quinquenal de idade usando a estrutura por grupo quinquenal de idade que presentan nas *Proyecciones de poboación a curto prazo* referidas á media anual.

3. PROYECCIONES DE TAXAS DE ACTIVIDADE NOS GRUPOS ELEMENTAIS

A partir das series homoxéneas de taxas de actividade anuais desde 1987, proxectamos as taxas de actividade nos grupos elementais utilizando modelos de extrapolación loxística. Para os grupos elementais masculinos proxectamos directamente as taxas de actividade mentres que para os grupos elementais femininos proxectamos as taxas de actividade indirectamente, xa que non proxectamos as taxas de actividade femininas senón as ratios entre as taxas de actividade femininas e masculinas.

Como xa adiantamos, as proxeccións das taxas de actividade nos grupos elementais permítenos calcular como agregación ponderada a proxección da taxa de actividade de calquera grupo poboacional que se obteña como unión de grupos elementais, sendo o coeficiente de ponderación de cada grupo elemental a proporción que representa nas proxeccións de poboación a súa poboación respecto á poboación total do grupo poboacional de interese.

3.1. PROYECCIONES DE TAXAS DE ACTIVIDADE MASCULINAS POR GRUPO QUINQUENAL DE IDADE

Se queremos extrapolar no tempo as taxas de actividade masculinas para o grupo quinquenal de idade i ($i \in \{16-19, 20-24, \dots, 65-69, 70+\}$) entre os valores asíntóticos inferior AI e superior AS resulta interesante considerar a transformación loxística:

$$TAT_{homes,i}^t = \ln \left(\frac{TA_{homes,i}^t - AI}{AS - TA_{homes,i}^t} \right)$$

xa que se modelamos a transformación loxística das taxas de actividade como a adición dunha tendencia temporal $f(t)$ e un residuo ε :

$$TAT_{homes,i}^t = f(t) + \varepsilon$$

temos que:

$$TA_{homes,i}^t = \frac{AI + AS \cdot e^{TAT_{homes,i}^t}}{1 + e^{TAT_{homes,i}^t}} = \frac{AI + AS \cdot e^{f(t)+\varepsilon}}{1 + e^{f(t)+\varepsilon}}$$

Así pois, se a partir das taxas de actividade masculinas para o grupo quinquenal de idade i no período de observación obtemos as correspondentes taxas de actividade transformadas loxisticamente e axustamos o modelo de tendencia temporal para a transformación loxística das taxas de actividade:

$$\widehat{TAT}_{homes,i}^t = \hat{f}(t)$$

podemos extrapolar no tempo as taxas de actividade para cada ano do período de proxección mediante:

$$\widehat{TA}_{homes,i}^t = \frac{AI + AS \cdot e^{\widehat{TAT}_{homes,i}^t}}{1 + e^{\widehat{TAT}_{homes,i}^t}} = \frac{AI + AS \cdot e^{\hat{f}(t)}}{1 + e^{\hat{f}(t)}}$$

Debemos ter en conta que existen tantos modelos de extrapolación loxística para as taxas de actividade masculinas para o grupo quinquenal de idade i como posibles valores asíntóticos para as taxas de actividade e tendencias temporais para a transformación loxística das taxas de actividade.

O procedemento considerado para obter as proxeccións de taxas de actividade masculinas para o grupo quinquenal de idade i consiste en:

1. Para o período de observación calculamos as taxas de actividade transformadas loxisticamente mediante tres transformacións loxísticas:

Transformación loxística 1. Simplemente garantimos que non se chegue a resultados absurdos, no sentido de que as proxeccións de taxas de actividade sempre queden acoutadas entre 0 e 1 por moi bruscas que sexan as tendencias ascendentes ou descendentes que se observaran no pasado:

$$AI = 0$$

$$AS = 1$$

Transformación loxística 2. Permitimos que as proxeccións de taxas de actividade queden por debaixo do mínimo histórico ou por encima do máximo histórico ata unha cantidade μ que representa a volatilidade absoluta media da serie histórica de taxas de actividade en períodos temporais de 15 anos, é dicir:

$$AI = \min\{TA_{homes,i}^t, t \text{ no período de observación}\} - \mu$$

$$AS = \max\{TA_{homes,i}^t, t \text{ no período de observación}\} + \mu$$

onde:

$$\mu = \frac{\sum_{t=t_{ini}}^{t_{fin}-15} |TA_{homes,i}^{t+15} - TA_{homes,i}^t|}{t_{fin} - t_{ini} - 14}$$

sendo t_{ini} e t_{fin} o primeiro e último ano do período de observación.

Transformación loxística 3. Asegurámonos de que as proxeccións de taxas de actividade sempre queden acoutadas entre o mínimo histórico e o máximo histórico:

$$AI = \min\{TA_{homes,i}^t, t \text{ no período de observación}\}$$

$$AS = \max\{TA_{homes,i}^t, t \text{ no período de observación}\}$$

2. Axustamos por mínimos cadrados ordinarios tres modelos de tendencia temporal para cada transformación loxística das taxas de actividade:

Tendencia temporal 1. As taxas de actividade transformadas loxisticamente son independentes do tempo:

$$f(t) = a$$

Tendencia temporal 2. As taxas de actividade transformadas loxisticamente presentan unha tendencia lineal no tempo:

$$f(t) = a + b \cdot t$$

Tendencia temporal 3. As taxas de actividade transformadas loxisticamente mostran unha tendencia cadrática no tempo:

$$f(t) = a + b \cdot t + c \cdot t^2$$

3. Utilizando as transformacións inversas asociadas ás transformacións loxísticas obtemos os valores axustados das taxas de actividade para o período de observación correspondentes aos nove modelos de extrapolación loxística axustados e quedámonos con aquel que presente un menor erro absoluto medio entre a serie axustada e a serie histórica nos últimos anos do período de observación.
4. Mediante o modelo de extrapolación loxística axustado que seleccionamos, extrapolamos a taxa de actividade para cada ano do período de proxección.

3.2. PROXECCIONS DE TAXAS DE ACTIVIDADE FEMININAS POR GRUPO QUINQUENAL DE IDADE

A ferramenta estatística utilizada para proxectar as taxas de actividade femininas tamén son os modelos de extrapolación loxística, pero en vez de empregalos para proxectar directamente as taxas de actividade femininas, usámoslos para proxectar as ratios de actividade entre mulleres e homes. A partir das proxeccións de ratios de actividade e taxas de actividade masculinas podemos obter as proxeccións de taxas de actividade femininas.

O procedemento utilizado para obter as proxeccións de taxas de actividade femininas para o grupo quinquenal de idade i ($i \in \{16-19, 20-24, \dots, 65-69, 70+\}$) é:

1. Para o período de observación calculamos as ratios de actividade entre mulleres e homes, é dicir, as ratios entre as taxas de actividade femininas e masculinas:

$$RA_i^t = \frac{TA_{mulleres,i}^t}{TA_{homes,i}^t}$$

e as transformamos loxisticamente mediante a transformación loxística asociada aos valores asíntóticos 0 e 1:

$$RAT_i^t = \ln \left(\frac{RA_i^t}{1 - RA_i^t} \right)$$

Só consideramos a transformación loxística asociada aos valores asíntóticos 0 e 1, pois nas últimas décadas estase producindo un proceso de converxencia das taxas de actividade femininas coas masculinas e se ben as ratios de actividade poden ser superiores a 1 non parece razoable esperar que as taxas de actividade femininas tendan a estabilizarse nun nivel superior ao correspondente ás taxas de actividade masculinas.

2. Axustamos por mínimos cadrados ordinarios tres modelos de tendencia temporal para a transformación loxística das ratios de actividade (constante: $f(t) = a$, lineal: $f(t) = a + b \cdot t$ e cadrática: $f(t) = a + b \cdot t + c \cdot t^2$).
3. Utilizando a transformación inversa asociada á transformación loxística obtemos os valores axustados das ratios de actividade para o período de observación correspondentes aos tres modelos de extrapolación loxística axustados e quedámonos con aquel que presente un menor erro absoluto medio entre as series axustada e histórica das ratios de actividade nos últimos anos do período de observación.
4. Mediante o modelo de extrapolación loxística axustado que seleccionamos, extrapolamos a ratio de actividade para cada ano do período de proxección:

$$\widehat{RA}_i^t = \frac{e^{\widehat{f}(t)}}{1 + e^{\widehat{f}(t)}}$$

5. Para cada ano do período de proxección obtemos a proxección da taxa de actividade como:

$$\widehat{TA}_{mulleres,i}^t = \widehat{RA}_i^t \cdot \widehat{TA}_{homes,i}^t$$

4. INDICADORES DERIVADOS DAS PROXECCIÓNS DE POBOACIÓN ACTIVA

A partir das proxeccións de poboación activa podemos proxectar o comportamento futuro dunha serie de indicadores que permiten analizar o mercado laboral desde diferentes perspectivas (sustentabilidade, feminidade e xuventude). Recordemos que a proxección da poboación activa de calquera grupo poboacional obtense multiplicando a súa taxa de actividade proxectada pola súa proxección de poboación.

4.1. ÍNDICES DE SUSTENTABILIDADE

O **índice de sustentabilidade dos maiores** definido como o cociente entre a poboación activa e a poboación inactiva de 65 ou máis anos:

$$\frac{\text{Activos}}{\text{Inactivos de 65 ou máis anos}}$$

proporciona información sobre a sustentabilidade do sistema público de pensións, xa que representa o número de potenciais traballadores por cada xubilado, entendendo o termo xubilado nun sentido amplo como toda persoa inactiva de 65 ou máis anos.

Agora ben, a poboación inactiva non só está formada polos xubilados senón tamén polos nenos, estudantes, incapacitados permanentes..., polo que para avaliar a sustentabilidade global dos sistemas públicos tamén resulta interesante o **índice de sustentabilidade global** definido como o cociente entre a poboación activa e a poboación inactiva:

$$\frac{\text{Activos}}{\text{Inactivos}}$$

4.2. ÍNDICE DE FEMINIDADE

Unha das características do mercado laboral nas últimas décadas é o progresivo acercamento das taxas de actividade femininas ás masculinas, polo que resulta interesante dispoñer dun indicador da estrutura por sexo da poboación activa.

O **índice de feminidade** definido como o cociente entre a poboación activa feminina e a poboación activa masculina:

$$\frac{\text{Mulleres activas}}{\text{Homes activos}}$$

sinala o número de mulleres activas por cada home activo.

4.3. ÍNDICE DE XUVENTUDE

A poboación está experimentando un proceso de envellecemento que ten o seu reflexo na poboación activa. Así pois, resulta interesante dispoñer dun indicador da estrutura por idade da poboación activa.

O **índice de xuventude** definido como o cociente entre a poboación activa entre os 16 e 39 anos e a poboación activa entre os 40 e 64 anos:

$$\frac{\text{Activos entre os 16 e 39 anos}}{\text{Activos entre os 40 e 64 anos}}$$

representa o número de activos entre os 16 e 39 anos por cada activo entre os 40 e 64 anos.

5. PRINCIPAIS RESULTADOS

A taxa de actividade da poboación de 16 ou máis anos pasa do 53,28% en 2018 ao 50,48% en 2032, o que supón unha caída de 2,80 puntos porcentuais. Agora ben, a redución que experimentará a taxa de actividade da poboación de 16 ou máis anos entre 2018 e 2032 non terá a mesma intensidade nas mulleres ca nos homes, xa que nos homes pasa do 57,52% en 2018 ao 53,07% en 2032 mentres que nas mulleres pasa do 49,38% en 2018 ao 48,13% en 2032.

En canto a poboación entre os 16 e 64 anos a taxa de actividade pasa do 73,44% en 2018 ao 74,60% en 2032, sendo diferente o comportamento para homes e mulleres, xa que para os homes baixará 0,79 puntos porcentuais mentres que para as mulleres subirá 3,10 puntos porcentuais.

Ano	Poboación de 16 ou máis anos			Poboación entre os 16 e 64 anos		
	Homes	Mulleres	Total	Homes	Mulleres	Total
1987	67,81	40,04	53,27	79,61	49,77	64,53
1988	66,84	41,36	53,50	78,31	50,93	64,48
1989	67,35	40,14	53,10	78,57	49,88	64,09
1990	66,23	39,85	52,41	77,92	50,05	63,85
1991	65,18	39,04	51,50	77,23	49,26	63,12
1992	63,82	39,40	51,05	75,90	49,74	62,71
1993	62,96	39,33	50,60	74,82	49,51	62,06
1994	62,07	39,21	50,12	74,24	49,83	61,95
1995	61,27	39,72	50,01	73,09	50,48	61,70
1996	61,18	38,80	49,47	73,46	49,95	61,62
1997	60,16	38,03	48,58	73,26	49,79	61,43
1998	60,30	38,24	48,75	74,03	50,70	62,27
1999	60,14	39,08	49,11	74,40	52,22	63,22
2000	60,87	40,32	50,11	75,43	54,05	64,65
2001	60,37	39,98	49,69	75,08	53,76	64,34
2002	60,79	40,95	50,39	75,99	55,34	65,58
2003	62,13	43,11	52,17	77,61	58,45	67,96
2004	62,48	44,75	53,20	78,00	60,71	69,29
2005	62,46	44,39	53,01	77,79	59,91	68,79
2006	61,81	45,71	53,39	77,30	62,06	69,63
2007	62,78	45,95	53,99	78,60	62,55	70,53
2008	63,18	47,03	54,76	79,35	64,05	71,67
2009	61,97	48,26	54,82	78,23	65,88	72,04
2010	61,21	47,75	54,19	77,83	65,62	71,71
2011	61,05	49,01	54,78	78,16	67,76	72,95
2012	61,01	49,50	55,01	78,42	68,68	73,54
2013	59,86	48,98	54,19	77,37	68,49	72,92
2014	58,98	49,01	53,78	76,79	69,02	72,89
2015	58,60	48,74	53,46	76,82	68,92	72,85
2016	58,48	48,84	53,45	77,04	69,39	73,19
2017	57,88	49,11	53,30	76,51	69,97	73,22
2018	57,52	49,38	53,28	76,32	70,59	73,44
2019	57,37	49,52	53,27	76,46	71,11	73,77
2020	57,15	49,66	53,24	76,52	71,58	74,03
2021	56,92	49,78	53,19	76,56	72,06	74,29
2022	56,54	49,74	52,99	76,41	72,31	74,35
2023	56,18	49,67	52,78	76,28	72,55	74,40
2024	55,80	49,57	52,54	76,13	72,72	74,41
2025	55,41	49,44	52,29	75,99	72,87	74,42
2026	55,03	49,30	52,04	75,87	73,02	74,44
2027	54,67	49,15	51,78	75,74	73,15	74,43
2028	54,33	48,99	51,54	75,63	73,28	74,44
2029	54,02	48,82	51,30	75,59	73,42	74,49
2030	53,71	48,62	51,05	75,58	73,54	74,55
2031	53,40	48,39	50,78	75,56	73,63	74,59
2032	53,07	48,13	50,48	75,53	73,69	74,60

Táboa 1: Taxas de actividade por sexo da poboación de 16 ou máis anos e entre os 16 e 64 anos

Nas figuras 1 e 2 podemos observar que tanto para os homes como para as mulleres o número de activos en 2032 será menor que en 2018, sendo esta baixada moito máis acusada nos homes ca nas mulleres.

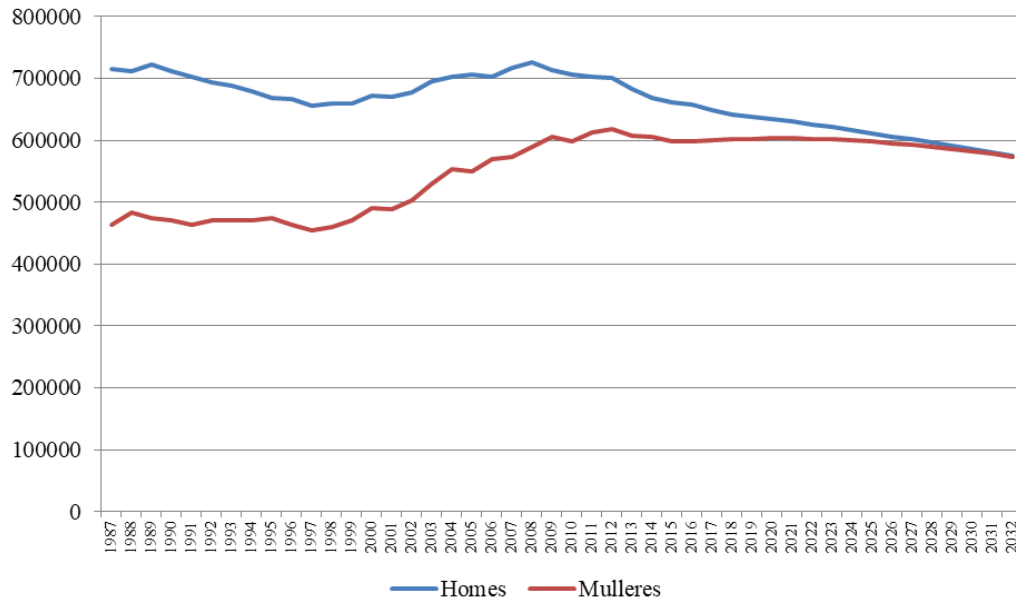


Figura 1: Poboación activa de 16 ou máis anos por sexo

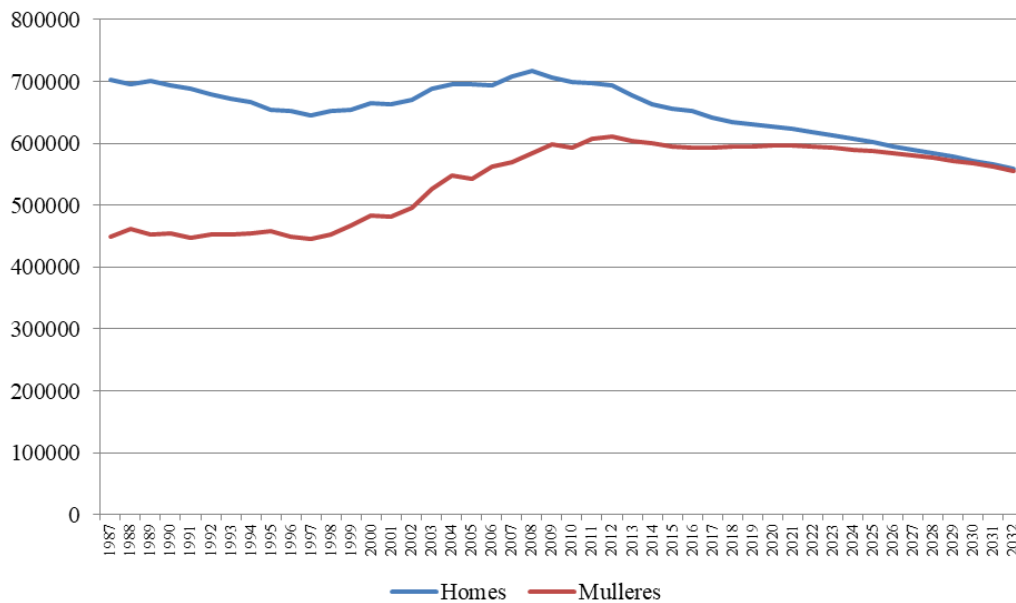


Figura 2: Poboación activa entre os 16 e 64 anos por sexo

Nas figuras 3 e 4 temos as curvas de actividade para homes e mulleres, onde podemos destacar que o grupo decenal de idade no que máis aumentará a taxa de actividade é o correspondente aos homes e mulleres entre os 55 e 64 anos, coherente co retraso que cabe esperar na idade efectiva de xubilación.

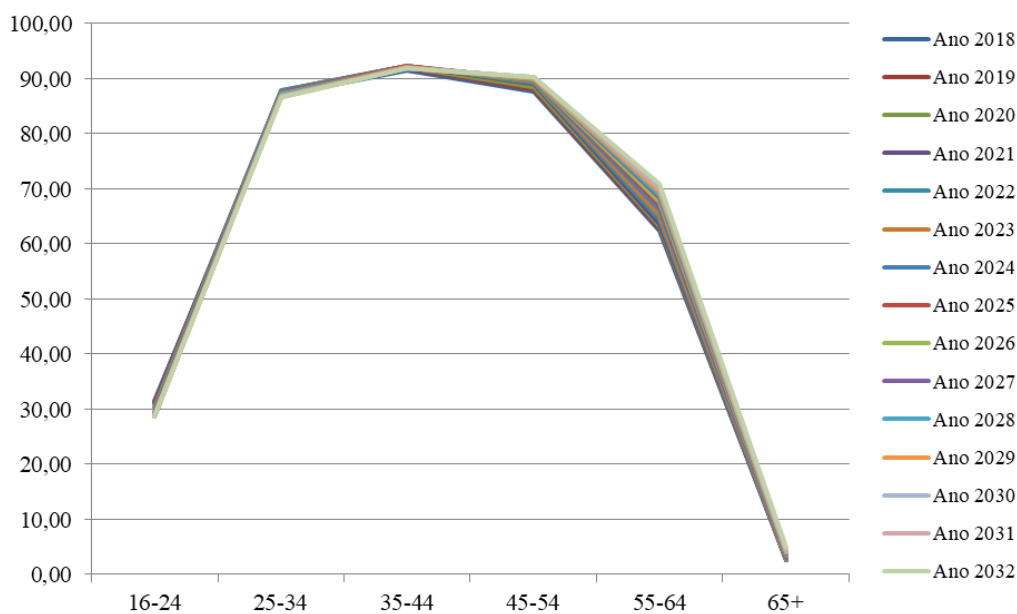


Figura 3: Curvas de actividade masculinas

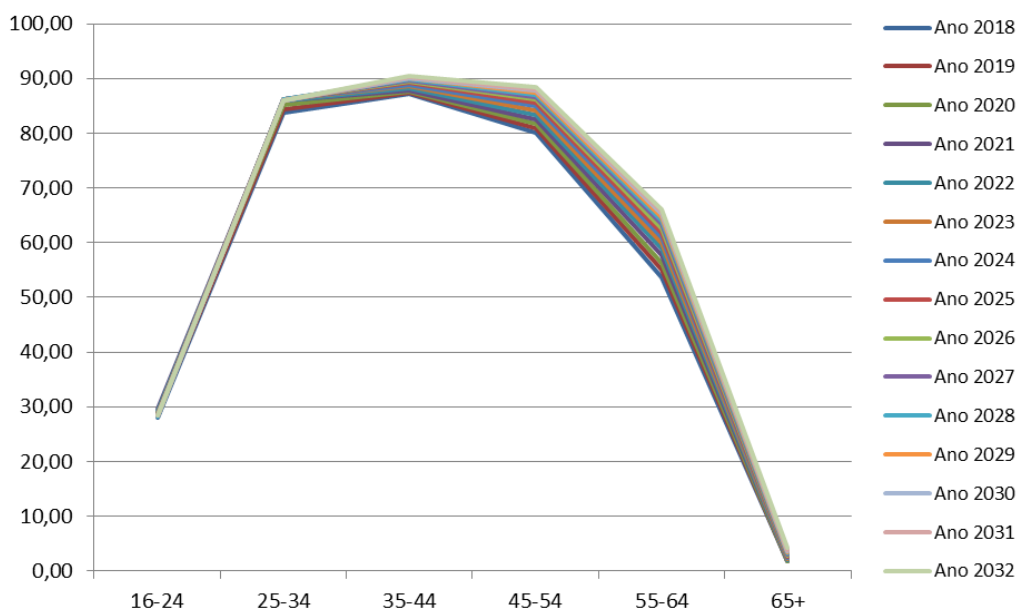


Figura 4: Curvas de actividade femininas

Os índices de sustentabilidade mostran que en 2032 reducirase o número de potenciais traballadores por cada inactivo, sobre todo por cada inactivo maior. Así mesmo, segundo os índices de feminidade e xuventude a poboación activa en 2032 será paritaria entre homes e mulleres pero máis vella.

Ano	Índice de sustentabilidade dos maiores	Índice de sustentabilidade global	Índice de feminidade	Índice de xuventude
2018	1,93	0,87	0,94	0,60
2019	1,90	0,87	0,94	0,58
2020	1,88	0,87	0,95	0,55
2021	1,86	0,87	0,96	0,54
2022	1,83	0,87	0,96	0,52
2023	1,80	0,86	0,97	0,50
2024	1,77	0,86	0,97	0,49
2025	1,74	0,85	0,98	0,49
2026	1,71	0,85	0,98	0,48
2027	1,68	0,85	0,99	0,48
2028	1,65	0,84	0,99	0,49
2029	1,63	0,84	0,99	0,49
2030	1,60	0,83	0,99	0,49
2031	1,57	0,83	1,00	0,50
2032	1,54	0,82	1,00	0,51

Táboa 2: Indicadores derivados das proxeccións de poboación activa

REFERENCIAS

- Instituto Nacional de Estadística (2016), Proyecciones de tasas de actividad globales y específicas por grupos de edad y sexo 2016-2029
- Instituto de Estadística de Andalucía (2011), Proyección de población activa de Andalucía 2009-2035
- Institut d'Estadística de Catalunya (2016), Projeccions de poboació activa 2015-2051
- Institut National de la Statistique et des Études Économiques (2017), Projections de la population active à l'horizon 2070
- Institute for Employment Research of the Federal Employment Agency (2017), Forecasting labour supply and population: An integrated stochastic model
- European Commission (2017), The 2018 ageing report: Underlying assumptions & Projection methodologies
- International Labour Office (2011), Projections of economically active population: A review of national and international methodologies
- International Labour Office (2017), Labour force estimates and projections 1990-2030
- International Monetary Fund (2018), Labor force participation in advanced economies: Drivers and prospects

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

ANÁLISE E MODELAXE DE OPINIÓNS SOBRE A CALIDADE DOS SERVIZOS PRESTADOS POLO SECTOR HOSTELEIRO OURENSÁN

Diego Frade Amil¹, Tomás R. Cotos-Yáñez², Manuel Mosquera-Rodríguez² e Ana
Pérez-González²

¹Técnico de Investigación, Dpto. Estatística e Investigación Operativa, Universidade de Vigo

²Dpto. Estatística e Investigación Operativa, Universidade de Vigo

RESUMO

Nos últimos anos, o uso de métodos estatísticos e de aprendizaxe automática orientados á resolución de problemas emerxentes na análise de datos no sector turístico acadou un notable crecemento. Este feito reflectiuse no aumento do índice de impacto que teñen revistas científicas como *Tourism Management*, *Annals of Tourism Research* ou *Journal of Travel Research*. Por outra banda, a aparición e aumento de plataformas web onde os usuarios poden valorar a súa experiencia durante a súa estada en establecementos hoteleiros, dotou ao sector dun mecanismo efectivo para recoller información de especial importancia tanto para os xestores dos establecementos como para os futuros usuarios. De feito, o 86% dos viaxeiros non reservan aloxamento sen antes ler as opinións dos usuarios sobre o mesmo (TripBarometer do ano 2017-2018). A finalidade principal deste traballo é o desenvolvemento de ferramentas para apoiar a calidade e mellora continua no sector hoteleiro ourensán baseadas no análise dos comentarios dos usuarios nas plataformas de TripAdvisor e Booking.com.

Palabras e frases chave: opinións sector hoteleiro, análise estatístico, machine learning, text mining.

1. INTRODUCCIÓN

Na actualidade, o “prestixio on-line” no sector hoteleiro resulta especialmente importante dado que incide notablemente nas reservas de aloxamento e está dispoñible para os potenciais clientes en servizos de reserva on-line (como Booking.com) ou portais de promoción turística (como Tripadvisor.es). A análise dos comentarios que fan os diferentes hóspedes permite ao empresario hoteleiro mellorar notablemente a clasificación (medida do prestixio on-line) do seu establecemento. O tratamento estatístico da información abrangue un amplo conxunto de posibilidades, dende resultados sinxelos (descritivos) ata modelos complexos de clasificación ou predición. A minería de datos atopa de forma automática ou semiautomática tendencias ou regras que explican o comportamento dos datos nun determinado contexto. O presente traballo xurde como resultado do subproxecto de investigación AMOCa: *Análise e Modelaxe de Opinións sobre a Calidade dos servizos prestados polo sector hoteleiro ourensán* dentro do proxecto global Desenvolvemento de ferramentas para apoiar a calidade e mellora continua no sector hoteleiro ourensán financiados pola Deputación de Ourense e convocados pola Universidade de Vigo para grupos de Investigación do Campus de Ourense no 2019.

Este proxecto pretende extraer datos de comentarios sobre hoteis da provincia de Ourense das plataformas Booking.com e Tripadvisor.es (dous importantes portais nos que se recadan as opinións dos turistas) e posteriormente obter información aplicando técnicas estatísticas e/ou de minería de datos.

2. OBXECTIVOS

Este subproxecto ten o seu inicio á finalización doutro subproxecto *EPOCa: Extracción e Pre-procesamento de Opinións sobre a Calidade dos servizos prestados polo sector hoteleiro ourensán* que fundamentalmente ten como obxectivo a extracción e pre-procesamento dos comentarios dos usuarios no sector hoteleiro ourensán nas plataformas TripAdvisor e Booking.com.

O obxectivo principal deste traballo é o resume e análise dos comentarios de hóspedes en establecementos hosteleiros da provincia ourensán aplicando técnicas estatísticas (Hastie *et al.* 2012) e de Data Mining (Tan *et al.* 2019) para a análise de textos (principalmente aprendizaxe supervisada e non supervisada).

Doutra banda, perseguimos obter perfís das mensaxes ou comentarios en función de diversas características que se poidan identificar nos mesmos. Esta tarefa realizarase a partir de matrices de proximidade en función de categorías para cada tema ou para as mensaxes dos usuarios, estableceranse clusters ou conglomerados de mensaxes/temas que permitirán a automatización de respostas (chatbots) simplificando a tarefa do xestor. Por exemplo, a identificación de mensaxes tipo: "Hotel xenial", "Todo perfecto" permite ao xestor contestar a este tipo de mensaxes cunha ou con varios tipos de resposta "estándar" predefinidas. Como resultado, permitirase agrupar mensaxes e/ou temas de comentarios dos hóspedes en función de diferentes características medidas nos mesmos, como presenza de determinadas palabras ou frases, valoración numérica da estancia, datos da estancia no hotel, ...

Finalmente avaliaremos a consistencia entre a valoración cuantitativa e cualitativa do establecemento para cada comentario. Os comentarios teñen unha descrición numérica (cuantitativa) que en moitos casos non está xustificada polos comentarios dos hóspedes (cualitativas). Este tipo de análise pode servir para identificar "fake text messages", por exemplo introducidos pola competencia. O uso de modelos lineais xeneralizados que explican a valoración cuantitativa en función da información recollida (principalmente información cualitativa), permite estimar e cuantificar o impacto de cada atributo sobre a puntuación total asignada. Este feito pode axudar ao xestor a definir estratexias de inversión para "elevar" a satisfacción dos seus clientes. O resultado será a identificación daqueles comentarios que presentan inconsistencias, incongruencias entre a valoración numérica e o propio comentario, cuantificar o aumento ou perda na valoración media, ao variar os diferentes atributos dos comentarios; en definitiva, atopar a mellor aproximación lineal que explique a valoración do establecemento en función dos comentarios.

3. RESULTADOS PRELIMINARES

Ata o de agora dispoñemos soamente dun conxunto de comentarios (inicial) de TripAdvisor. Como resumo, na Taboa 1 e na Figura 1 amósanse o número de establecementos hosteleiros e a súa distribución espacial por concello (dereita da Figura 1) para a plataforma TripAdvisor, destacar que dos 92 concellos ourensáns soamente 56 teñen establecementos rexistrados en TripAdvisor.

As variables recollidas en TripAdvisor poden agruparse do seguinte xeito:

- Valoracións dos servizos: valoración-total, valoración da ubicación, valoración da limpeza, valoración do servizo, valoración calidade-prezo e valoración individual.
- Información dos servizos ofertados polo establecemento hosteleiro de tipo binario (153 variables), por exemplo: se ten restaurante, zona infantil, aparcadoiro gratuito, wifi gratuito, piscina, ...

Ademáis dispoñemos de información adicional recollida do Directorio de empresas e establecementos turísticos (Turismo de Galicia 2019), como por exemplo a posición xeográfica, o número de camas, ...

Datos Tripadvisor	
Comentarios realizados	5947
Usuarios con comentarios	5478
Establecementos hosteleiros	165
Concellos	56

Táboa 1: Resumo dos datos totais obtidos de TripAdvisor.

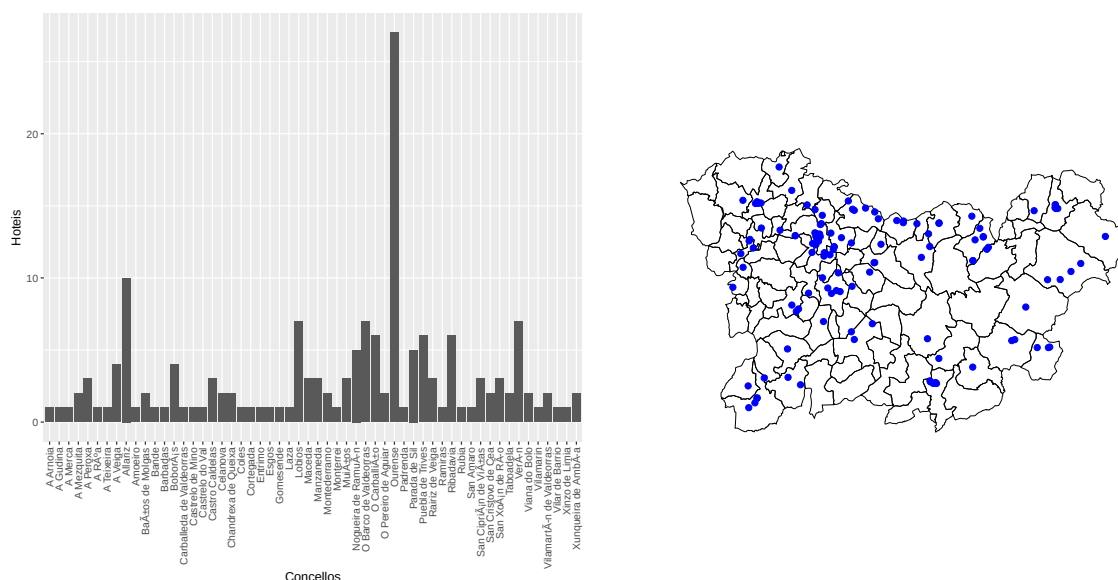


Figura 1: Número (esquerda) e distribución espacial (dereita) de establecementos hosteleiros por concello (extraído de TripAdvisor).

AGRADECEMENTOS

O traballo dos autores foi apoiado parcialmente pola subvención INOU19-04B da Universidade de Vigo e a Deputación de Ourense.

REFERENCIAS

- Hastie T., Tibshirani R. e Friedman J. (2012). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.
- Tan P.N., Steinbach, M., Karpadne, A., e Kumar, V. (2019). Introduction to data mining. Pearson.
- Tripadvisor (2019) TripBarometer 2017-2018. Global report. <https://www.tripadvisor.es/TripAdvisorInsights/w4594>
- Turismo de Galicia (2019) Directorio de empresas e establecementos turísticos. https://www.turismo.gal/espazo-profesional/directorio-de-empresas-e-establecementos-turisticos?langId=es_ES

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

UNHA APROXIMACIÓN Á AVALIACIÓN DO IMPACTO RECENTE DE FONDOS EUROPEOS EN GALICIA

Veiguela Fernández, N¹, López Vizcaino ME¹, Del Río Viqueira, I¹, Iglesias Patiño, CL¹, Jácome Rodríguez, R¹, Calvo Ocampo, ME¹, Barros Torres, A², Martínez Barral, JM², Landeira Caamaño, J², Andión Hermida, E³, Tarrío Tobar, A³

¹ Instituto Galego de Estatística

² Consellería de Medio Rural

³ Consellería de Facenda

RESUMO

Os Fondos Europeos son instrumentos públicos importantes para apoiar accións relacionadas coas prioridades clave da Unión Europea (UE), que se canalizan a través de programas plurianuais nos Estados membros e nas rexións. Polo tanto, redunda no interese dos responsables políticos e do público saber se o diñeiro produce os efectos esperados para acadar os obxectivos políticos da UE. Neste relatorio preséntanse dúas propostas de avaliación do impacto recente en Galicia dos fondos europeos do Programa de Desenvolvemento Rural 2014-2020 e do Programa Operativo FEDER Galicia 2014-2020.

Palabras e frases chave: avaliación de impacto, fondos europeos, regresión loxística, contrafactual, mostraxe.

1. INTRODUCCIÓN

Os Fondos son instrumentos públicos importantes para apoiar accións relacionadas coas prioridades clave da Unión Europea (UE), que se canalizan a través de programas plurianuais nos Estados membros e nas rexións. Polo tanto, redunda no interese dos responsables políticos e do público saber se o diñeiro se gasta de maneira responsable, produce os efectos esperados e se as intervencións se dirixen aos beneficiarios adecuados para acadar os obxectivos políticos da UE (Comisión Europea, 2018). A avaliación é un instrumento para valorar a eficacia, a eficiencia, os resultados e os impactos da política, a pertinencia, a coherencia e o valor engadido da UE a fin de garantir a súa responsabilidade e transparencia, e, en última instancia, mellorar a súa concepción. Este traballo céntrase na avaliación dos impactos destes fondos.

Neste punto xorde a pregunta ¿como verificar que os cambios observados se deben ás intervencións e que enfoques de avaliación deben utilizarse a tal fin?

A avaliación de impacto mide os resultados da actuación pública illando os efectos da intervención do resto de efectos que puidesen incidir no resultado. O gran reto da avaliación de impacto é determinar que parte do resultado obtido se debe á intervención pública, illándoo doutros fenómenos que poden afectar á situación que a intervención quere solucionar. En definitiva, deséxase establecer a causalidade entre a intervención e o resultado obtido. Para iso é necesario establecer un escenario válido que permita comparar o que ocorreu despois da intervención co que ocorrería se a intervención non se levase a cabo. Como observar dous estados da natureza simultaneamente non é posible, a clave da avaliación de impacto está na construción de contrafactuais válidos (European Network for Rural Development, 2018).

Existen diversas alternativas á hora de realizar unha avaliación de impacto dunha política pública; estas pódense clasificar en dous grupos: os métodos en deseños experimentais (experimentos aleatorios) e todos os demais métodos, que se denominan case-experimentais. A diferenza entre ambos reside en que nos deseños experimentais se parte dunha poboación potencial de beneficiarios da actuación a avaliar, e os individuos participan ou non de acordo cun mecanismo de asignación aleatorio. Os individuos non participantes, que se denominan “grupo de control”, constitúen o contrafactual. Os deseños case experimentais diferéncianse dos experimentais en que a participación non se decide mediante un sistema

aleatorio, senón que se determina ou ben de forma voluntaria por parte dos participantes, ou porque outro axente decide esa participación conforme a uns criterios preestablecidos, ou por ambas as razóns, sendo o contrafactual o grupo (ou un subgrupo) de non participantes. Neste traballo empregaremos deseños case-experimentais.

A análise contrafactual debe basearse en comparacións entre grupos de beneficiarios e non beneficiarios do programa, que sexan o máis similares posibles (en dimensións observables e non observables). Se os dous grupos son semellantes desde o punto de vista estatístico, pódese supor que calquera diferenza nos resultados débese ao programa, xa que os efectos externos actuarán de igual forma sobre ambos os grupos.

A construción dun grupo de control (de non beneficiarios) pode ser unha tarefa esixente debido a unha sólida autoselección do programa e as estritas condicións de concesión das subvencións. Os beneficiarios do programa que recibiron apoio, poden diferir en gran medida dos non beneficiarios, tanto en termos de características estruturais como de rendemento económico, medioambiental e socioeconómico. Neste punto botarase man de técnicas estatísticas para resolver o problema que se plantea, técnicas que axudarán a superar moitas das dificultades que xorden durante o proceso.

Este relatorio céntrase na avaliación do impacto de dous fondos: o Programa de Desenvolvemento Rural (PDR) de Galicia 2014-2020 financiado polo Fondo Europeo Agrícola de Desenvolvemento Rural (FEADER) e os fondos do Programa Operativo FEDER Galicia 2014-2020.

O PDR de Galicia foi adoptado oficialmente pola Comisión Europea o 18 de novembro de 2015, e nel preséntanse as prioridades de Galicia no que respecta á utilización dos 1.186,4 millóns de euros de fondos públicos dispoñibles para o septenio 2014-2020. Este novo programa dará continuidade á política en materia de desenvolvemento rural que vén realizando a Xunta de Galicia co apoio dos fondos comunitarios do Fondo Europeo Agrícola de Desenvolvemento Rural (FEADER), máis concretamente a través do Programa de Desenvolvemento Rural 2007-2013, que finalizou a súa vixencia a finais de 2015. Neste período de financiamento comunitario 2014-2020, Galicia percibirá 889,8 millóns de euros do FEADER no marco da política agraria comunitaria, supoñendo un desembolso público de 1.186,4 millóns de euros no rural galego a través do PDR 2014-2020, dos cales a Xunta de Galicia alcanzará un importe superior a 217 millóns de euros, mentres que a contribución do Ministerio de Agricultura, Alimentación e Medio Ambiente será de 79 millóns de euros.

O PDR 2014-2020 persigue 6 prioridades:

- Prioridade 1: Fomentar a transferencia de coñecementos e a innovación na agricultura, a silvicultura e as zonas rurais
- Prioridade 2: Viabilidade das explotacións, competitividade e xestión sustentable dos bosques
- Prioridade 3: Fomentar a organización da cadea alimentaria, incluíndo a transformación e a comercialización dos produtos agrícolas, o benestar animal e a xestión de riscos
- Prioridade 4: Restaurar, conservar e mellorar os ecosistemas relacionados coa agricultura e a silvicultura
- Prioridade 5: Eficiencia dos recursos e paso a unha economía hipocarbónica e adaptable aos cambios climáticos nos sectores agrícola, alimentario e forestal
- Prioridade 6 : A inclusión social e o desenvolvemento local, promovendo a redución da pobreza e a inclusión social e o desenvolvemento económico nas zonas rurais

Tendo en conta as prioridades anteriores, o PDR de Galicia 2014-2020 establece tres liñas básicas de actuación en concordancia cos obxectivos definidos pola Política Agraria Común (PAC):

- Fomentar a competitividade da agricultura galega mediante a modernización das explotacións agrícolas e forestais, a incorporación de novos agricultores en explotacións orientadas ao mercado e con capacidade de diversificación económica, a ordenación do territorio e a mellora da produtividade da industria agroalimentaria e forestal promovendo a produción de alto valor engadido.
- Conservar e revalorizar o patrimonio natural de Galicia, promover a xestión sustentable dos recursos naturais e a loita contra o cambio climático
- Mellorar a calidade de vida da poboación rural fomentando a prestación de servizos básicos que axuden á fixación da poboación no medio rural, a creación de emprego e a redución do risco de pobreza

O PDR fai un especial fincapé no relevo xeracional no agro galego. Destinará 66,8 millóns de euros para apoiar os mozos e mozas que queiran instalarse para desenvolver actividades agrícolas. A mellora das explotacións contará cun apoio de máis de 166 millóns de euros. Estas liñas de apoio están acompañadas co apoio ás pequenas explotacións con 14,4 millóns de euros. A formación de asesores, o emprego de servizos e asesoramento e o establecemento de servizos de substitución axudarán á mellora da rendibilidade e conciliación da vida laboral e familiar das familias dedicadas á produción de alimentos deste país, e investiranse preto de 17 millóns de euros. Ademais, para a mellora da base territorial e dos accesos ás explotacións investiranse 57,4 millóns de euros.

O programa Operativo FEDER Galicia 2014-2020 promove para Galicia un cambio de modelo económico orientado cara os factores de crecemento e competitividade. En liña cos obxectivos da Europa 2020 e a estratexia de especialización intelixente rexional, o programa trata de impulsar o desenvolvemento económico sustentable na rexión mediante a creación de emprego –en particular nos sectores de alto valor engadido- e a mellora da competitividade da economía rexional co apoio ás PEMES.

O custo total previsto do Programa Operativo FEDER Galicia, actualizado en 2017, ascende a 1.142,1 millóns de euros, entre recursos públicos e privados, dos que o 80% corresponde á Axuda comunitaria (913,7 millóns de euros), o 16,3% corresponde á cofinanciación da Xunta de Galicia (186,5 millóns de euros), e o 3,7% restante (42 millóns de euros) será achegado polo sector privado. Este importe distribúese a través de 9 obxectivos temáticos ou eixos e 17 prioridades de investimento (Táboa 1).

Táboa 1 Importes programados por obxectivo temático ou eixe do Programa Operativo FEDER Galicia. Período 2014-2020.

Obxectivos temáticos	Total
OT1. Potenciar a investigación, o desenvolvemento tecnolóxico e a innovación	178.000
OT2. Mellorar o uso e calidade das TIC e o acceso ás mesmas	175.875
OT3. Mellorar a competitividade das PEME, do sector agrícola (no caso do FEADER) e do sector da pesca e a acuicultura (no caso do FEMP)	178.750
OT4. Favorecer a transición a unha economía baixa en carbono en todos os sectores	229.032
OT5. Promover a adaptación ao cambio climático e a prevención e xestión de riscos	49.423
OT6. Conservar e Protexer o medio e promover a eficiencia dos recursos	190.068
OT7. Promover o transporte sostible e eliminar os estrangulamentos nas infraestruturas de rede fundamentais	0
OT8. Promover a sostibilidade e a calidade no emprego e favorecer a mobilidade laboral	0
OT9. Promover a inclusión social e loitar contra a pobreza e calquer forma de discriminación	64.298
OT10. Investir en educación, formación e formación profesional para a adquisición de capacidades e un aprendizaxe permanente	59.163
OT11. Mellorar a capacidade institucional das autoridades públicas e as partes interesadas e a eficiencia da administración pública	0
OT13. Asistencia técnica	17.501
TOTAL	1.142.110

Datos en miles de euros

Nota: non se destinan Fondos FEDER aos obxectivos temáticos 7,8 e 11.

As actuacións que se cofinanciarán co FEDER Galicia 2014-2020 están relacionadas fundamentalmente co impulso das seguintes prioridades estratéxicas: a investigación e a innovación; as tecnoloxías da información e a comunicación; a competitividade das PEMEs; unha economía baixa en carbono; a adaptación ao cambio climático e a prevención de riscos; a protección ambiental e a

promoción da eficiencia de recursos; a promoción da inclusión social e a loita contra a pobreza e a discriminación; e o investimento en educación e formación.

Nas seguintes seccións expónse a metodoloxía seguida para avaliar o impacto do PDR e do Programa Operativo FEDER en Galicia.

2. METODOLOXÍA PARA A AVALIACIÓN DO IMPACTO DO PDR 2014-2020

O obxectivo deste punto é establecer a metodoloxía que permita avaliar o posible impacto das axudas do PDR 2014-2020, no referido á medida 4 do PDR (investimentos en activos físicos: explotacións agrícolas), en aspectos referidos a súa dimensión económica e produtiva. Por ser a medida 4 do actual programa, semellante no seu alcance á medida 121 do PDR 2007-2013 (modernización de explotacións), neste estudo inténtase recoller a evolución das explotacións de produción de leite que foron subvencionadas durante o período 2007-2017 pola Xunta de Galicia e con cargo a fondos FEADER dos PDR. A elección do sector produtivo de vacún de leite vén motivada pola importancia deste sector en Galicia (peso económico e social na economía galega).

Analizarase a evolución das explotacións tendo en conta as variables de produción de leite, do número de vacas de muxidura, do número de empregados e da renda agraria (medida a nivel de explotación e por traballador). Para avaliar o desempeño das explotacións subvencionadas seleccionáranse un grupo de explotacións que non recibiron subvención (grupo control) e comparáronse cos resultados das explotacións subvencionadas. Designaremos ao grupo das explotacións subvencionadas como “grupo obxectivo”.

Dispónse dunha base de datos proporcionada pola Consellería de Medio Rural coas explotacións de produción de leite subvencionadas no período 2007-2017, coas características destas explotacións no momento de pedir a subvención. Nesta base de datos tamén se dispón das explotacións que durante ese período solicitaron axudas do PDR pero non se lles concedeu en ningún ano do período de estudo (grupo candidato a formar parte do grupo control).

A modo de contexto, traballouse cos datos do período 2007-2017 e considerouse que para poder realizar unha análise do posible efecto do PDR, é necesario contar cunha serie anual ampla que permita extraer conclusións. Unha vez detallada a información para a serie completa tamén se fará unha análise partindo do ano 2015 (ano cero para o PDR 2014-2020). Polo tanto, se se considera que a actual medida 4 do PDR de Galicia 2014-2020, é similar á medida 121 do PDR 2007-2013, este estudo é en realidade unha avaliación ex-post do período precedente que permite aventurar o alcance do actual programa.

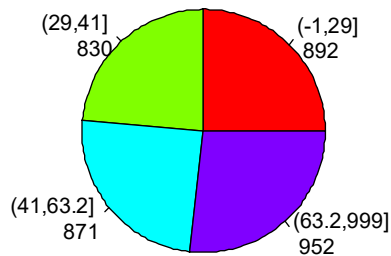
2.1 DETERMINACIÓN DO GRUPO CONTROL

O grupo candidato a control está formado por explotacións de produción de leite que durante o período 2007-2017 solicitaron axudas ao investimento, pero non foron beneficiarios. De todos os xeitos, o grupo de control debería ter características similares ao grupo obxectivo, polo tanto, de todas as explotacións proporcionadas que durante o período 2007-2017 solicitaron axudas ao investimento, pero non a conseguiron, seleccionáronse aquelas que eran máis similares ás do grupo obxectivo atendendo ao seguinte procedemento:

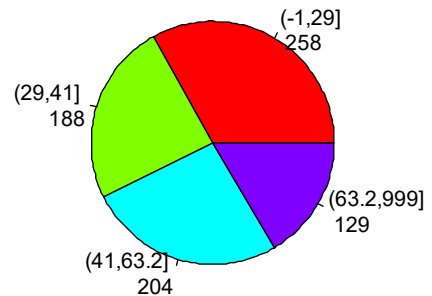
- Determinouse a distribución das explotacións de produción de leite segundo o número de vacas e a superficie no grupo obxectivo e no grupo inicial candidato a ser grupo control. Neste paso comprobouse que as explotacións do grupo candidato a control eran máis pequenas cás do grupo obxectivo, tal e como se mostra nos seguintes gráficos

Gráfico 1.- Distribución inicial do número de vacas e da superficie agraria no grupo obxectivo e no grupo candidato a control

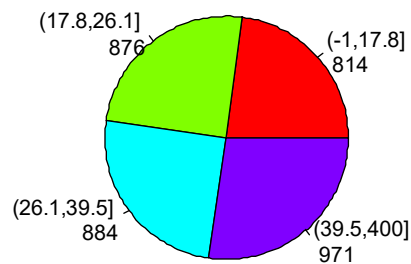
Nº de vacas. Grupo obxectivo



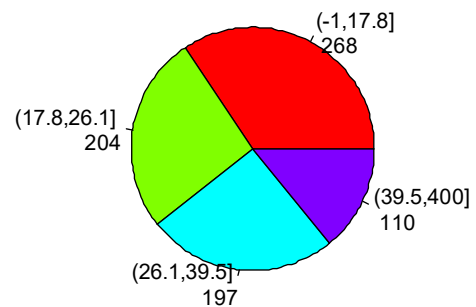
Nº de Vacas. Grupo candidato a control



Superficie. Grupo obxectivo



Superficie. Grupo candidato a control



- Empregouse un modelo de regresión loxística (Silva, 1994) e calculouse a probabilidade de que cada explotación de produción de leite do grupo candidato a control puidese pertencer ao grupo obxectivo, e elimináronse aquelas explotacións que teñen unha probabilidade máis baixa de pertencer ao grupo obxectivo. O modelo de regresión loxística empregado foi:

$$\Pr(\text{tratamento} = 1) = \frac{\exp(\beta_0 + \beta_1 \text{superficie} + \beta_2 \text{leite} + \beta_3 \text{vacas}/\text{superficie})}{1 + \exp(\beta_0 + \beta_1 \text{superficie} + \beta_2 \text{leite} + \beta_3 \text{vacas}/\text{superficie})}$$

onde

tratamento=1 se a explotación pertence ao grupo obxectivo e *tratamento*=0 se a explotación pertence ao grupo candidato a control

superficie: hectáreas da explotación, proporcionadas no momento de solicitar á subvención

leite: produción de leite en litros no momento de solicitar a subvención

vacas: nº de vacas no momento de solicitar a subvención

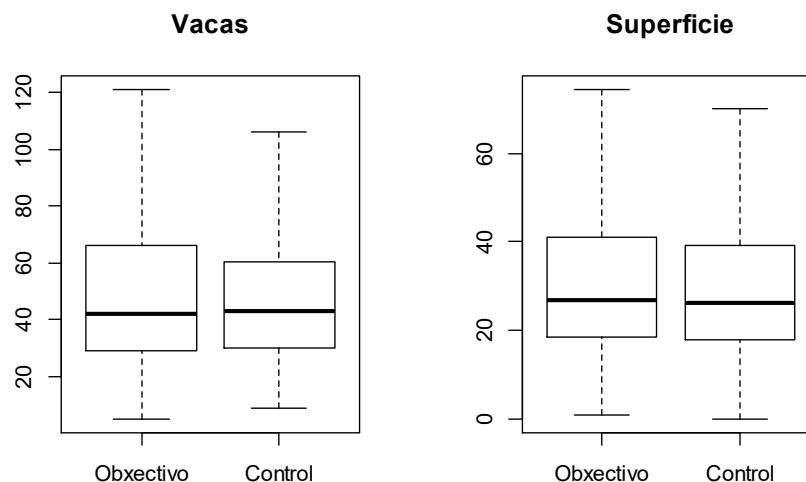
$\beta_0, \beta_1, \beta_2, \beta_3$: coeficientes da regresión

Logo de determinar o modelo calculouse a probabilidade de cada explotación de produción de leite do grupo candidato a control de pertencer ao grupo obxectivo (o propensity score). Elimináronse as explotacións que tiñan unha probabilidade inferior a 0,75.

- Visualizouse a distribución segundo o número de vacas e a superficie das explotacións do paso anterior para o grupo obxectivo e o grupo candidato a control e comprobouse que aínda persistían as diferenzas entre o grupo obxectivo e o grupo candidato a control.
- Tendo en conta os resultados do paso anterior decidiuse extraer unha mostra estratificada (Tillé *et al.*, 2016) no grupo candidato a control para tratar de equilibrar as distribucións das dúas variables (superficie e nº de vacas) no grupo obxectivo e no grupo control. Extraeuse unha mostra de 100 explotacións no grupo candidato a control para cada un dos estratos considerados na distribución da superficie, [0;17,8], (17,8;26,1], (26,1;39,5] e (39,5;999]. Estes estratos fórmanse a partir dos cuartís da distribución da superficie de todas as explotacións de produción de leite. Con esta mostra formada por 400 explotacións conseguiuase equilibrar a distribución.
- Empregouse o test Chi2 para comprobar a similitude das dúas distribucións. Cunha significación do 95% comprobouse que a distribución da superficie e do número de vacas non eran significativamente diferentes no grupo obxectivo e no grupo control ($p=0,7$ e $p=0,056$, respectivamente)

Nas seguintes gráficas pódese visualizar os diagramas de caixas que comparan as distribucións da superficie e do número de vacas nos dous grupos

Gráfico 2.- Diagrama de caixas coa distribución da superficie agraria e o número de vacas das explotacións dos grupos obxectivo e de control



Nota: nesta gráfica, para unha mellor visualización das distribucións, elimináronse os valores atípicos.

2.2 SEGUIMIENTO DAS EXPLOTACIÓNS DO GRUPO OBXECTIVO E DO GRUPO CONTROL

Logo de determinar as explotacións que forman o grupo obxectivo e o grupo control cómpre avaliar a evolución destes grupos tendo en conta catro variables: o número de vacas de muxidura, a produción de leite, o número de empregados e a renda agraria.

As fontes de información son:

- Produción de leite para o período 2007-2018, proporcionada polo Fondo Galego de Garantía Agraria (FOGGA)
- Superficie agraria das explotacións para o período 2007-2018, proporcionada polo FOGGA

- Renda Agraria para o período 2007-2017, proporcionada pola Agencia Tributaria (AT) a partir dos ficheiros de “Nivel de renda agraria do ano ...” que a Consellería de Medio Rural descarga (coa correspondente autorización) da sede electrónica da AT.
- Rexistro de bovino para o período 2007-2018, dispoñible no IGE e proporcionado pola Consellería de Medio Rural.
- Rexistro de Explotacións Agrarias de Galicia (REAGA), proporcionado pola Consellería de Medio Rural.
- Afiliacións á Seguridade Social e Contas de Cotización, dispoñible no IGE e proporcionados pola Tesorería General de la Seguridad Social.
- Red Contable Agraria Nacional (RECAN), proporcionada polo Ministerio de Agricultura, Pesca y Alimentación.

Todas as fontes de información, a excepción da RECAN, proporcionan datos a nivel de explotación agraria identificada. Neste traballo utilizarase como identificador o CIF da explotación, o que permitirá cruzar unhas fontes con outras.

O primeiro paso é construír unha base de datos onde estarán todas as explotacións do grupo obxectivo e do grupo control, para o período dende 2007 ata o 2018, coa información sobre o número de vacas de muxidura, a produción de leite, a superficie, a renda agraria e o número de traballadores.

Para a construción desta base de datos atopáronse limitacións que se detallan a continuación, xunto coa solución adoptada en cada caso

- **Vacas de muxidura:** en ausencia de datos do número de vacas en muxidura para algunha das explotacións incluídas na base de datos de estudo, para algún ano en concreto
 - o Calculouse a produtividade media por vaca en cada explotación ao longo do período dispoñible
 - o No ano no que non estaba dispoñible o número de vacas asignóuselle un valor tendo en conta a produción de leite dese ano e a produtividade media por vaca desa explotación.
- **Renda agraria:** Dependendo dos anos, dispónse de información de varios tipos:
 - o Período 2010-2012: dispónse de información proporcionada pola AT para as explotacións que pediron as axudas neste período e para os dous anos anteriores. Os datos faltantes neste período, predicíronse empregando un modelo de regresión lineal para cada ano. Este modelo ten como variable resposta a renda agraria das explotacións das que se dispón deste dato e como variables explicativas o número de vacas de muxidura, superficie e a produción de leite, coa seguinte expresión:

$$Renda\ agraria_t = \beta_{0t} + \beta_{1t}superficie_t + \beta_{2t}leite_t + \beta_{3t}vacas_t + e_t$$

onde

$Renda\ agraria_t$: renda agraria da explotación en cada ano t

$superficie_t$: hectáreas da explotación en cada ano t

$leite_t$: produción de leite en litros en cada ano t

$Vacas_t$: nº de vacas de muxidura en cada ano t

$\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}$: coeficientes da regresión para cada ano t

e_t : erro da regresión que segue unha distribución normal $N(0, \sigma_t^2)$

Todos os modelos de regresión empregados para o período 2010-2012 teñen un R^2 superior a 0,6.

- o Período 2007-2009: a información dispoñible para as explotacións deste anos procedente da AT é moi incompleta. Por isto, empregouse para predicir os datos da renda agraria neste período o modelo de regresión lineal establecido no punto anterior para o ano 2010.
- o Período 2013-2017: empregamos os datos da AT proporcionados para todo este período e para todas as explotacións do grupo obxectivo e do grupo control. Neste punto tamén se solicitou a AT información sobre o percibido, por rendas das sociedades, do socio solicitante da axuda. A problemática que teñen estes datos é que a AT ten información sobre a participación do socio na renda da sociedade, pero descoñécese a porcentaxe de

participación de cada socio, que permita calcular a renda total da sociedade. Neste situación cruzáronse os ficheiros da AT cos do REAGA, que dispoñen da porcentaxe de participación de cada socio na sociedade. Para completar a información de todos os suxeitos contemplados na mostra, empregouse un modelo de regresión lineal similar ao do período 2010-2012.

Por último, fíxose unha avaliación da coherencia dos datos da renda agraria para todo o período 2008-2017 cos datos do número de vacas de muxidura, superficie, produción de leite e o ano de seguimento, axustando o seguinte modelo de regresión lineal

$$Renda\ agraria = \beta_0 + \beta_1 superficie + \beta_2 leite + \beta_3 vacas + \beta_4 t + e$$

onde

Renda agraria: renda agraria da explotación

superficie: hectáreas de superficie da explotación

leite: produción de leite en litros

vacas: nº de vacas de muxidura

t: ano de seguimento da explotación, toma valores do 2008 ao 2017

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$: coeficientes da regresión

e: erro da regresión, que segue unha distribución normal $N(0, \sigma^2)$

Neste modelo final, todas as variables explicativas son significativas cun nivel de confianza do 95% ($p < 0,05$) e o $R^2 = 0,7$.

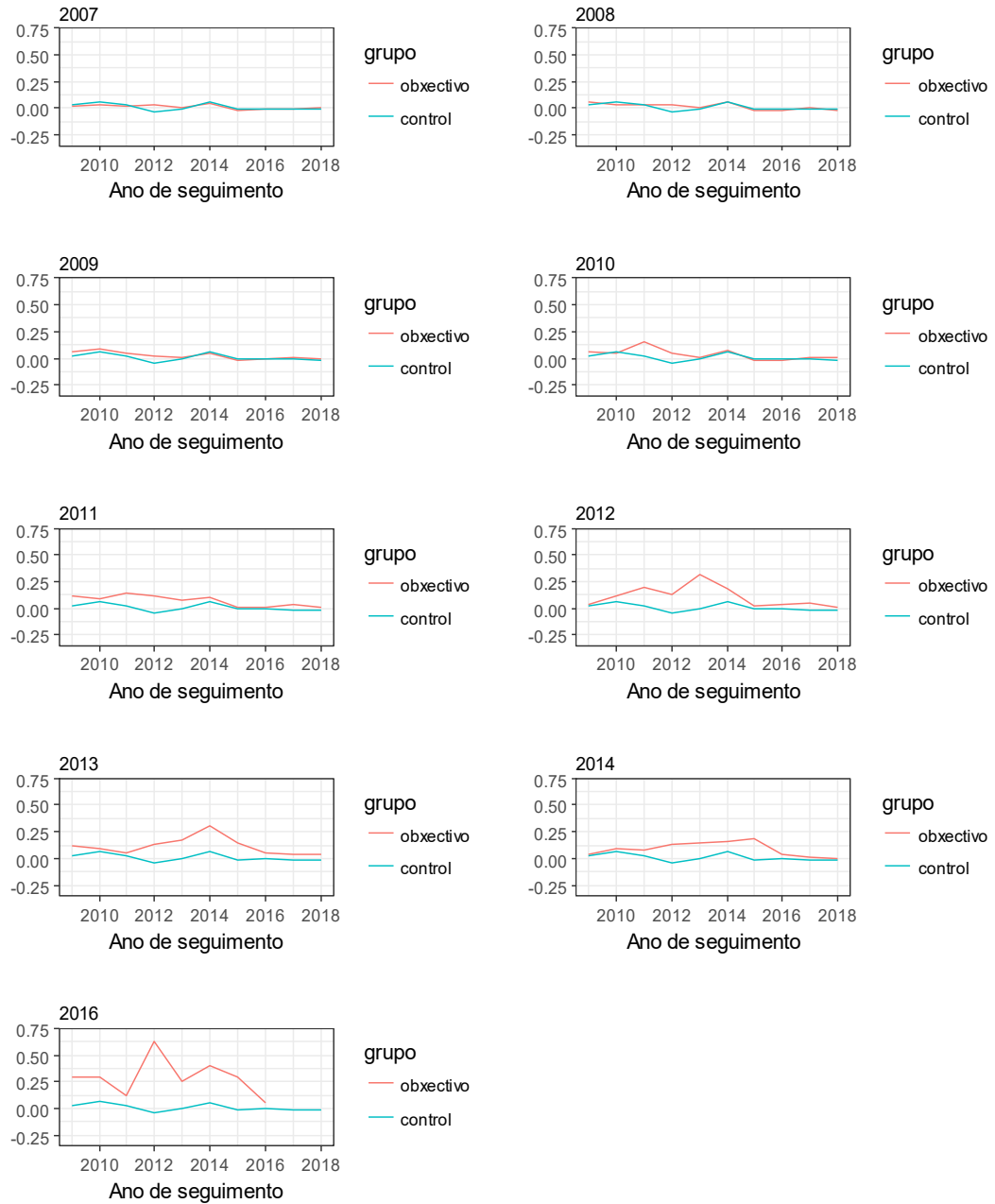
Unha vez que coñecemos que o axuste é aceptable avaliáronse os residuos. Para os casos nos que os residuos fosen demasiado grandes (10 casos), fíxose a predición de novo da renda agraria empregando este último modelo e imputouse con este resultado.

2.3 PRIMEIROS RESULTADOS

No período de estudo (2007-2017) cóntase con información de explotacións de produción de leite que recibiron as subvencións en diferentes momentos de tempo. Por simplicidade, clasificamos a cada unha das explotacións por data de resolución. Para os casos nos que unha mesma explotación tivese máis dunha resolución de axuda considérase o primeiro ano de resolución da axuda.

No seguinte gráfico preséntase a evolución das taxas de variación interanual do número de vacas de muxidura para as explotacións do grupo control e o grupo obxectivo, tendo en conta o ano de resolución da subvención:

Gráfico 3.- Taxas de variación anual (TVA) do número de vacas de muxidura por ano de resolución da subvención, para o grupo obxectivo e o grupo control



Para tratar de condensar toda a información que nos achegan as representacións do gráfico 3 nun só gráfico, o que se fixo foi calcular as taxas medias de variación da magnitude correspondente a 1, 2, 3,... anos, da data de resolución da axuda. Isto permítenos ter unha información máis homoxénea e resumida. Estas taxas calcularanse do seguinte xeito:

$$tv_1 = \frac{m_1 \sqrt[n_1]{(1 + tva_{2008-2007})^{n_{2007}} (1 + tva_{2009-2008})^{n_{2008}} \dots (1 + tva_{2018-2017})^{n_{2017}} - 1}}{m_2 \sqrt[n_2]{(1 + tva_{2009-2008})^{n_{2007}} (1 + tva_{2010-2008})^{n_{2008}} \dots (1 + tva_{2018-2016})^{n_{2016}} - 1}}$$

$$\dots$$

$$m_1 = n_{2007} + n_{2008} + \dots + n_{2016} + n_{2017}$$

$$m_2 = n_{2007} + n_{2008} + \dots + n_{2016}$$

onde

n_t : número de explotacións no ano t

tv_1 sería a taxa a 1 ano, tv_2 a taxa a 2 anos e así sucesivamente.

$tva_{t-t'}$: a taxa de variación interanual da magnitude empregada entre o ano t e o ano t' en tanto por 1

As fórmulas presentadas permiten comparar conxuntamente todos os datos dispoñibles dos grupos obxectivo e control, sen necesidade de facer distinción polo ano de resolución da subvención, xa que nestas a evolución temporal faise depender do número de períodos transcorridos dende a resolución da axuda.

Nos seguintes gráficos móstrase a representación destas taxas para o caso do número de vacas de muxidura, número de traballadores, produción de leite e renda agraria.

Gráfico 4.- Taxas medias de variación a 1, 2,... anos da data de resolución das axudas, do número de vacas de muxidura e da produción de leite, para o grupo obxectivo e o grupo control

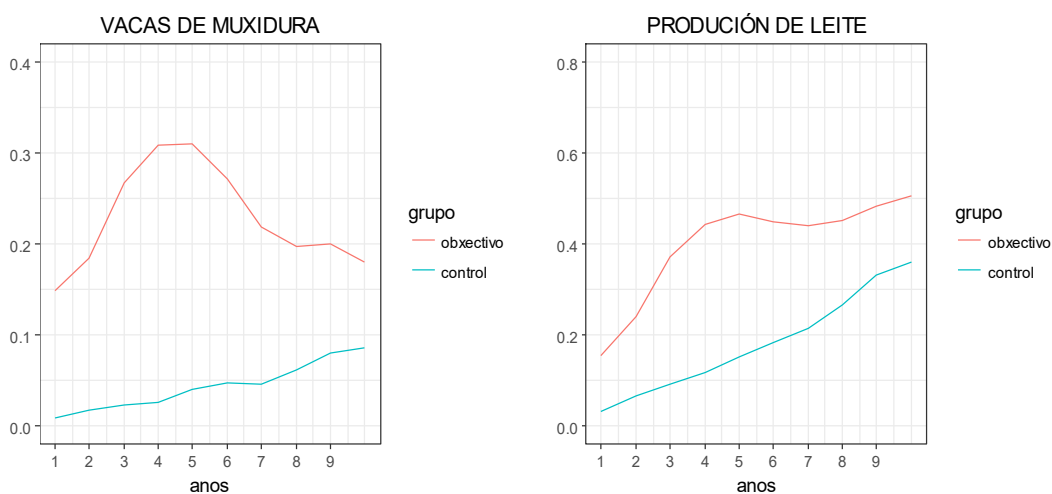
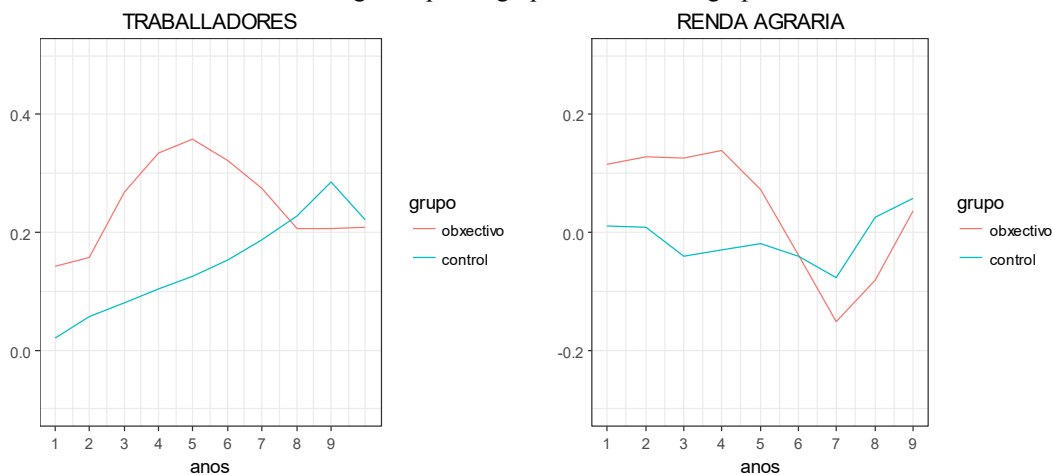


Gráfico 5.- Taxas medias de variación a 1, 2,... anos da data de resolución das axudas, do número de traballadores e da renda agraria, para o grupo obxectivo e o grupo control



Nota: Os datos da renda agraria non están deflactados

En todos os gráficos anteriores obsérvase un aparente impacto positivo das axudas ao investimento do PDR nas explotacións dos 2 aos 4 anos da data de resolución da subvención, este impacto parece que vai decrecendo unha vez pasados eses anos. No caso do número de vacas, da produción de leite ou do

número de traballadores, o impacto parece durar ata os 5 anos. Hai que ter en conta que estes gráficos poden estar influenciados pola problemática da desaparición das explotacións.

3. METODOLOXÍA PARA A AVALIACIÓN DO IMPACTO DO PROGRAMA OPERATIVO FEDER GALICIA 2014-2020

O obxectivo deste punto é avaliar o efecto da axuda do Programa Operativo FEDER Galicia 2014-2020 nas empresas apoiadas na Comunidade Autónoma de Galicia. Este efecto mídese seguindo a evolución económica (en termos de emprego e cifra de negocio) das empresas subvencionadas en 2016 con cargo a fondos do Programa Operativo FEDER Galicia. A análise restrínxese ás actuacións enmarcadas dentro dos obxectivos temáticos 1 (“potenciar a investigación, o desenvolvemento tecnolóxico e a innovación”) e 3 (“mellorar a competitividade das PEMES”), concretamente a través das liñas de actuación establecidas nos seguintes obxectivos específicos (OE):

- Obxectivo Específico 01.02.01: impulso e promoción de actividades de I+i lideradas polas empresas, apoio á creación e consolidación de empresas innovadoras e apoio á compra pública innovadora; con actuacións consistentes en axudas para fomentar a cooperación entre PEMES e os demais axentes do sistema rexional de innovación e axudas para a mellora da capacidade de innovación das empresas.
- Obxectivo Específico 03.01.02: creación de novas empresas e viveiros de empresas, en particular mellorando o acceso á financiación e a servizos de apoio avanzados; as actuacións consisten en axudas para proxectos de novos emprendedores.
- Obxectivo Específico 03.04.01: promover o crecemento e consolidación das PEMES, en particular mellorando a súa financiación, tecnoloxía e acceso a servizos de apoio avanzados, incluíndo os sectores agrícola, pesqueiro, mariño, marítimo, turístico, cultural, comercial e de rehabilitación de edificación, así como as PEMES e autónomos/as dedicados ao comercio ao retallo ou venda ambulante; as actuacións consisten en axudas ás empresas para financiar proxectos de investimento empresarial.
- Obxectivo Específico 03.04.03: promover a internacionalización das PEMES, con actuacións consistentes en axudas para aumentar a presenza en mercados exteriores.

Para avaliar o desempeño das empresas subvencionadas nos 4 obxectivos específicos comentados, establecerase un grupo de empresas (grupo control) que non recibiron subvención, similar ao subvencionado (grupo obxectivo), e compararanse os resultados económicos de ambos grupos.

3.1 DETERMINACIÓN DO GRUPO CONTROL

Para establecer o grupo control, empregouse un procedemento de emparellamento realizado mediante a librería MatchIt (Ho *et al.*, 2011) do software libre R (R Core Team, 2018). A técnica estatística que está detrás desta librería é a análise de propensións (máis coñecida polo seu nome en inglés propensity score analysis) (Fraile, 2014). Mediante esta técnica estímase uns índices (de propensión) para cada observación (neste caso, empresa integrante do marco) por medio dunha regresión loxística; cada un destes índices correspondería coa probabilidade que tería cada caso de pertencer ao grupo de tratamento (obxectivo) en base as súas características observadas. Unha vez estimados os índices de propensión, selecciónanse para o grupo control os casos cuxos índices se asemellen aos do grupo de tratamento (Ho *et al.*, 2011). Neste caso cada empresa do grupo obxectivo ten unha empresa no grupo control, a diferenza da análise que se fixo no PDR.

No caso que nos ocupa, as características observables incluídas na regresión loxística para emparellar ás empresas foron a condición xurídica, o subgrao de urbanización (IGE, 2016) do concello no que se atopan emprazadas, a rama de actividade e o emprego total de ambas a 31/12/2016; o método empregado para a selección de casos foi o emparellamento polo veciño máis próximo. O marco de empresas a partir do cal extraer o grupo control proporcionouno o ficheiro de contas de cotización da Seguridade Social. Tamén foi esta a fonte empregada para determinar as características das empresas do grupo obxectivo.

O procedemento de emparellamento para a busca do grupo control realizouse de forma iterativa e separada para cada obxectivo específico (OE), permitindo que unha mesma empresa puidese actuar de control de varias empresas obxectivo, sempre e cando se trate de grupos OE distintos; para cada tipo de subvención, realizáronse os pasos seguintes:

- 1ª Iteración: buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica (*condición xurídica*), a rama de actividade a 3 díxitos da CNAE-2009 (*rama 3 díx.*), a subzona de urbanización (*subgrao urbanización*) e a diferenza no emprego entre ambas fose inferior a 5 persoas (*diferenza emprego*). Ademais, para as empresas subvencionadas en base ao obxectivo específico 03.01.02 esixiuse que a empresa control fose de recente creación (medido pola variable data de *alta* na Seguridade Social). Neste paso conseguíuse emparellar a 305 empresas obxectivo con empresas do grupo control.
- 2ª Iteración: buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica agrupada, a rama a 2 díxitos, a subzona de urbanización e a diferenza de emprego entre ambas fose inferior a 5 persoas. Igual que no paso anterior, no caso da subvención 03.01.02 esixiuse que a empresa control fose de recente creación.
- 3ª Iteración: buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica agrupada, a rama a 2 díxitos e a diferenza de emprego entre ambas fose inferior a 5 persoas; é dicir, nesta iteración prescínlese da similitude no subgrao de urbanización.
- 4ª Iteración: buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica agrupada, a sección de actividade, a subzona de urbanización e a diferenza de emprego entre ambas fose inferior a 5 persoas; nesta iteración agrúpanse as ramas de actividade a 2 díxitos en seccións, seguindo a CNAE-2009 (IGE, 2009).
- 5ª Iteración: buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica agrupada, a sección de actividade e a diferenza de emprego entre ambas fose inferior a 5 persoas.
- 6ª Iteración: agrupouse de novo a condición xurídica da empresa en tipoloxías máis xenéricas e a sección de actividade en sectores (primario, industria, construción e servizos). Buscouse unha empresa control para cada empresa obxectivo de forma que en ambas coincidiran a condición xurídica agrupada, o sector e a diferenza de emprego entre ambas fose inferior a 5 persoas.
- 7ª Iteración: finalmente, para completar o grupo control replicouse o emparellamento realizado na sexta iteración (agrupación 2 condición xurídica e sector) pero, neste caso, relaxouse o criterio aplicado ao emprego: escolléronse empresas control cuxo volume de emprego se atopase entre o 80 e o 120% do volume de emprego das empresas obxectivo.

Unha vez determinado o emparellamento en cada paso do procedemento iterativo comentado é preciso corroborar que este se realizou correctamente. Para isto adóitase analizar se se conseguiu o equilibrio nas covariables (características observadas) de ambos grupos, obxectivo e control. A medición do dito equilibrio lévase a cabo mediante diferenzas estandarizadas e, aínda que non existe un criterio establecido respecto de onde situar este punto de equilibrio, sóese fixar nunha diferenza non superior a 20 puntos (Fraile, 2014). Neste estudo corroborouse que logo do proceso de emparellamento realizado en cada iteración se cumpría este requisito.

Ademais das corroboracións realizadas despois de cada iteración, unha vez completado o grupo control procedeuse á construción de táboas comparativas da estrutura de ambos grupos segundo as características analizadas e o obxectivo específico considerado (táboa 2). Existen algunhas disparidades menores se consideramos os niveis iniciais destas variables, por exemplo, a condición xurídica sen agrupar ou a rama a 3 díxitos da CNAE-2009. Pero a estrutura é case idéntica ao empregar a agrupación 1 e 2 da condición xurídica e o nivel sección da clasificación de actividades. Só se aprecian diferenzas importantes ao agrupar as empresas de cada grupo segundo o subgrao de urbanización do concello no que se emprazan. De feito, non se recomenda realizar a comparativa da evolución do grupo obxectivo e control por subgrao de urbanización para os obxectivos específicos que contén cun menor número de empresas; non obstante, si sería posible efectuar esta comparativa para a totalidade de ambos grupos, sen diferenciar o tipo de subvención.

Táboa 2. Estrutura dos grupos obxectivo (obxect.) e control (contr.) para cada obxectivo específico ao considerar distintas variables dos mesmos

Condición xurídica:

Obxectivo Específico	Autónomos/as			Sociedades			Outros tipos		
	Obxect.	Contr.	%	Obxect.	Contr.	%	Obxect.	Contr.	%
01.02.01	0	0	100	34	34	100	0	0	100
03.01.02	12	12	100	20	20	100	0	0	100
03.04.01	0	0	100	80	80	100	0	0	100
03.04.03	4	4	100	423	423	100	26	26	100
Total	16	16	100	557	557	100	26	26	100

Sector de actividade:

Obxectivo Específico	Primario			Industria		
	Obxect.	Contr.	%	Obxect.	Contr.	%
01.02.01	0	0	100	11	11	100
03.01.02	1	1	100	2	2	100
03.04.01	0	0	100	56	55	101,8
03.04.03	9	9	100	213	212	100,5
Total	10	10	100	282	280	100,7

Obxectivo Específico	Construción			Servizos		
	Obxect.	Contr.	%	Obxect.	Contr.	%
01.02.01	3	3	100	20	20	100
03.01.02	1	1	100	28	28	100
03.04.01	6	6	100	18	19	94,7
03.04.03	17	17	100	214	215	99,5
Total	27	27	100	280	282	99,3

Subgrao de urbanización:

Obxectivo Específico	ZDP			ZIP alta			ZIP baixa		
	Obx.	Con.	%	Obx.	Con.	%	Obx.	Con.	%
01.02.01	19	21	90,5	7	6	116,7	3	2	150,0
03.01.02	13	15	86,7	10	8	125,0	1	2	50,0
03.04.01	16	21	76,2	30	31	96,8	5	2	250,0
03.04.03	184	188	97,9	120	132	90,9	42	35	120,0
Total	232	245	94,7	167	177	94,4	51	41	124,4

Obxectivo Específico	ZPP alta			ZPP intermedia			ZPP baixa		
	Obx.	Con.	%	Obx.	Con.	%	Obx.	Con.	%
01.02.01	3	2	150,0	0	0	100	2	3	66,7
03.01.02	1	0	Inf.	3	4	75,0	4	3	133,3
03.04.01	10	7	142,9	2	0	Inf.	17	19	89,5
03.04.03	34	31	109,7	11	8	137,5	61	59	103,4
Total	48	40	120,0	16	12	133,3	84	84	100

A última variable que nos permite comparar a estrutura de ambos grupos é o volume de emprego en 2016. Dado que se trata dunha característica continua, podemos construír as funcións de densidade correspondentes a ambos grupos e comprobar se estas se asemellan. Da observación dos gráficos 6 e 7 dedúcese que o solapamento das funcións de densidade é substancial, tanto se os analizamos

conxuntamente ou considerando o tipo de subvención recibida. Tamén se pode concluír o mesmo se construímos as funcións de densidade do emprego agrupando as empresas en base as súas características (condición xurídica, sector de actividade e subgrao de urbanización).

Gráfico 6. Estimación tipo núcleo das funcións de densidade do emprego en 2016 das empresas obxectivo e control

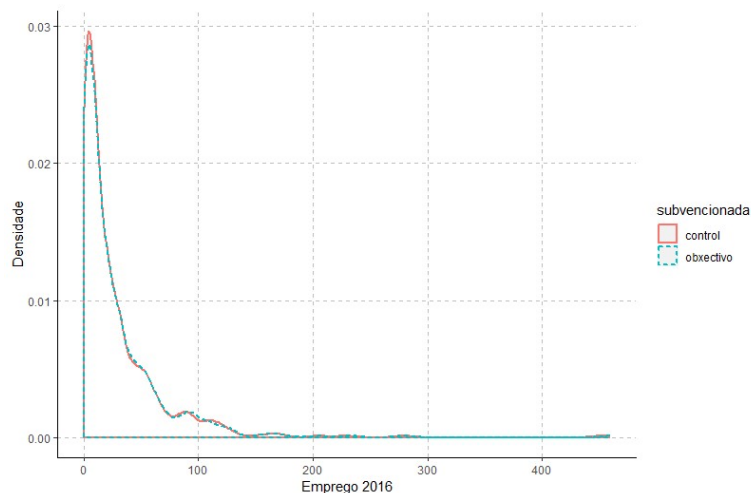
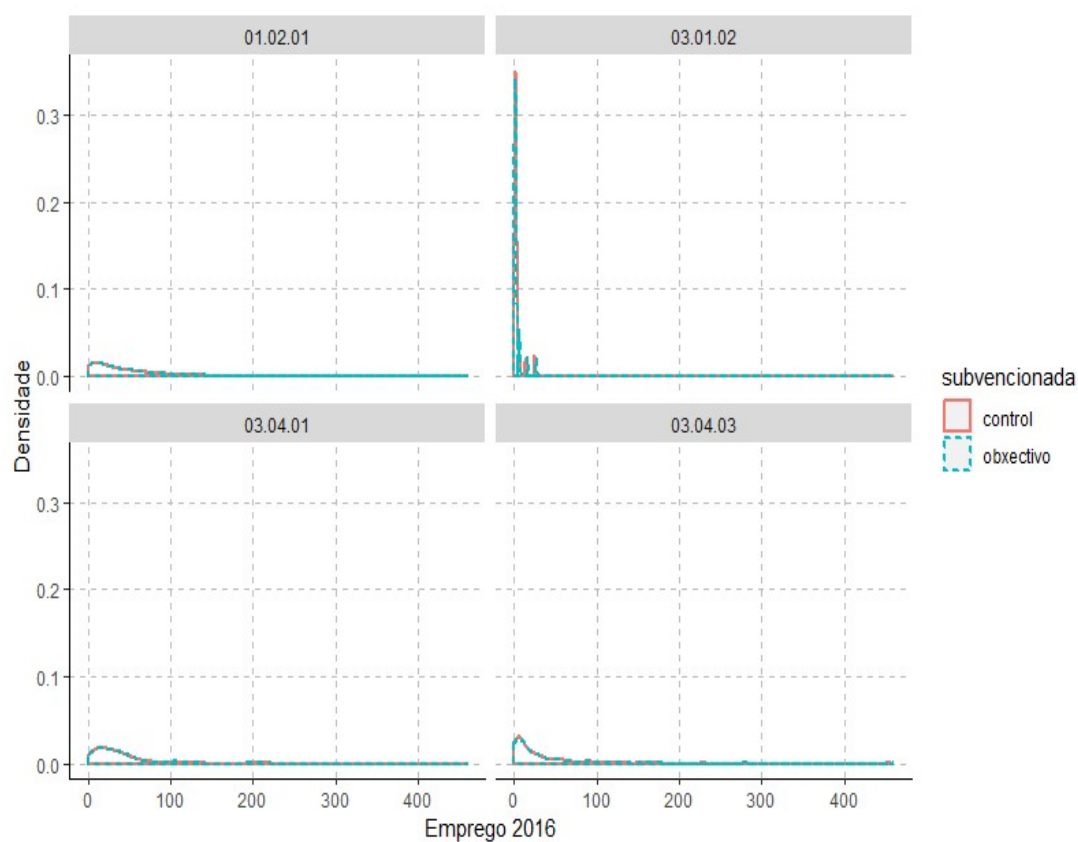


Gráfico 7. Estimación tipo núcleo das funcións de densidade do emprego en 2016 das empresas obxectivo e control agrupándoas en base ao tipo de subvención recibida (obxectivo específico)

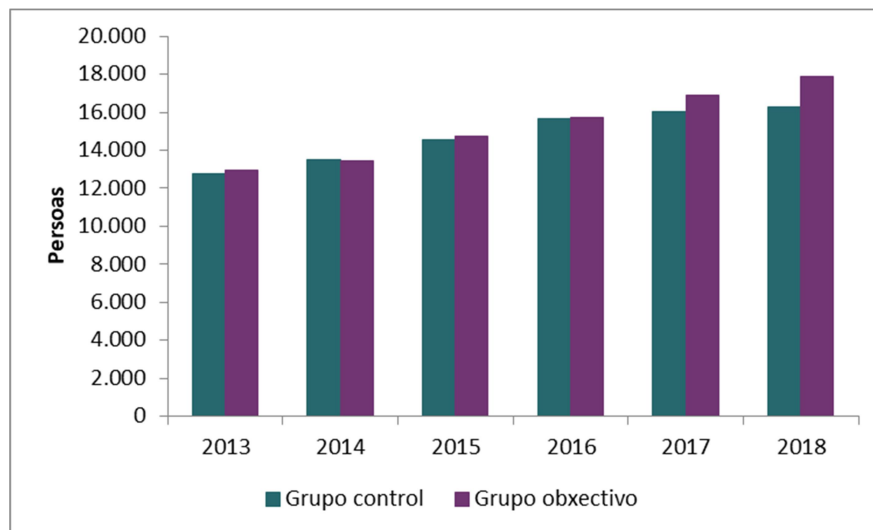


3.1 PRIMEIROS RESULTADOS

Os gráficos 8 e 9 comparan o volume de emprego (en Galicia) do grupo de empresas subvencionado e non subvencionado, tanto en termos absolutos como relativos.

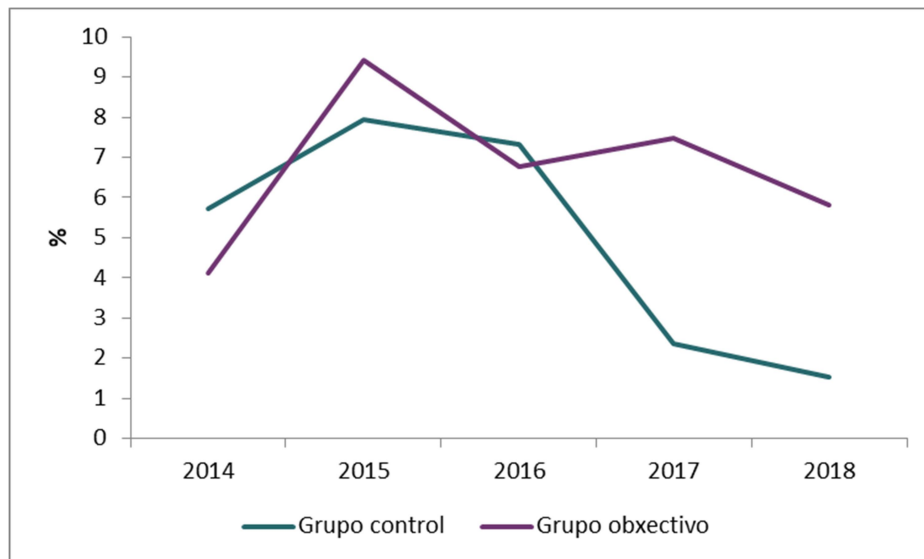
A evolución do emprego foi máis favorable para o grupo obxectivo que para o control, en particular tras a percepción da subvención: entre 2013 e 2018 o emprego das empresas subvencionadas pasou de 12.937 traballadores/as a 17.894; é dicir, aumentou en 4.957 persoas, o que en termos relativos se traduce nunha suba do 38,3%. O emprego das empresas non subvencionadas aumentou nunha contía inferior, tanto en termos absolutos (unhas 3.487 persoas), como relativos (suba do 27,3% entre 2013-2018). Se restrinximos a análise aos dous últimos anos do período de observación, 2017 e 2018, xusto despois de que se percibise a subvención, as discrepancias son incluso máis notorias: o grupo obxectivo experimentou taxas de variación interanuais do emprego do 7,5% e 5,8% respectivamente, namentres que no grupo control as taxas de variación foron moi inferiores en ambos exercicios: do 2,4% entre 2016-2017 e do 1,5% entre 2017-2018. Como se desprende tamén destes gráficos, ambos grupos parten de niveis de emprego similares (o do grupo obxectivo é lixeiramente superior ao do grupo control), o que resulta evidente se se ten en conta que esta variable foi incluída no proceso de emparellamento que se realizou para obter o grupo control¹.

Gráfico 8. Comparación da evolución do emprego das empresas subvencionadas (grupo obxectivo) e non subvencionadas (grupo control). Número de traballadores e traballadoras.



¹ A cifra de emprego obtívose para todos os anos en estudo do ficheiro de contas de cotización á Seguridade Social, ao que igual que xa se fixera coa variable “emprego” utilizada no proceso de emparellamento. Esta cifra fai referencia ao volume de traballo contratado por conta allea en exclusiva, sen incluír outras figuras contratuais que poidan ter relación coa empresa (por exemplo, non inclúe aos traballadores/as contratados/as a través dunha ETT, nin tampouco ás persoas que traballan en exclusiva para unha soa empresa baixo a forma de traballo autónomo). O emprego anual obtense como a media dos catro trimestres do ano; agora ben, os datos de cada trimestre están referidos ao último día do mesmo (31/marzo, 30/xuño, 30/setembro e 31/décembro). Se nun trimestre concreto unha empresa do grupo obxectivo ou control non figura no ficheiro correspondente, considérase que nese trimestre o emprego da empresa é 0 (as empresas só están obrigadas a manter aberta unha conta de cotización na nosa Comunidade Autónoma se contan con emprego en Galicia ou prevén a contratación dalgún traballador/a nunha data próxima á de extracción do ficheiro).

Gráfico 9. Comparación da evolución do emprego das empresas subvencionadas (grupo obxectivo) e non subvencionadas (grupo control). Taxa de variación interanual do emprego (%).



CONCLUSIÓNS

Neste relatorio preséntanse dúas propostas de avaliación dunha medida do PDR e de varios obxectivos específicos do Programa Operativo FEDER en Galicia. As propostas son diferentes porque dependen da información estatística da que se dispón para levar a cabo estas avaliacións. No caso do PDR foi necesario empregar un grupo control procedente de explotacións que pediron apoio pero non se lle concedeu, pola necesidade de dispoñer de información tributaria, e no caso do FEDER púidose seleccionar un grupo control dunha base de datos dispoñible no IGE. Trátase de dúas propostas diferentes, que se poden estender a outras medidas destes e doutros fondos.

Despois do traballo é necesario por en valor a axuda que proporcionaron as diversas técnicas estatísticas empregadas para resolver un problema real de avaliación de impacto de Fondos.

REFERENCIAS

Comisión Europea (Agosto 2018). DIRECTRICES. Valoración de los logros e impactos de los PDR en 2019. https://enrd.ec.europa.eu/evaluation/publications/assessing-rdp-achievements-and-impacts-2019_en

European Network for Rural Development. Thematic Working Group 6 - Data for the assessment of RDP achievements and impacts. https://enrd.ec.europa.eu/evaluation/thematic-working-groups/thematic-working-group-6-data-assessment-rdp-achievements-and_en?pk_campaign=TWG6&pk_source=ealert

Fraile Valderrey, A. (2014), “Línea de investigación: análisis estadístico con R; análisis de propensiones en R”, traballo fin de máster, Departamento de Estadística e Investigación Operativa da Universidad de Granada

Ho, D.E., Imai, K., King, G., Stuart, E. A. (2011), “MatchIt: Nonparametric Preprocessing for Parametric Causal Inference”, *Journal of Statistical Software*, volume 42 (xuño de 2011)

IGE (2016) Clasificación dos concellos segundo o grao de urbanización (GU 2016) http://www.ige.eu/web/mostrar_paxina.jsp?paxina=003003001&idioma=gl

IGE (2009). Clasificación Nacional de Actividades Económicas (CNAE-2009)
http://www.ige.eu/web/mostrar_paxina.jsp?paxina=003003&idioma=gl

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Silva, L.C. (1994). *Excursión a la regresión logística en ciencias de la salud*. Ediciones Díaz de Santos. S.A.

Tillé, Y. and Matei, A. (2016). sampling: Survey Sampling. R package version 2.8. <https://CRAN.R-project.org/package=sampling>

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

MODELIZACIÓN DA DEGRADACIÓN DE BIOMATERIAIS USADOS EN REXENERACIÓN ÓSEA

Salvador Naya¹, Javier Tarrío Saavedra², Yaroslava Robles³ e Silvia Díaz Prado⁴

¹Departamento de Matemáticas. Grupo MODES, ITMATI e CITIC. Escola Politécnica Superior. Universidade da Coruña.

²Departamento de Matemáticas. Grupo MODES e CITIC. Escola Politécnica Superior. Universidade da Coruña.

³Grupo de Investigación en Materiales (GiMaT). Universidad Politécnica Salesiana. Cuenca. Ecuador.

⁴Instituto de Investigación Biomédica de A Coruña (INIBIC). Departamento de Fisioterapia, Medicina e Ciencias Biomédicas. Universidade da Coruña.

RESUMO

Neste traballo presentamos distintos modelos estatísticos usados para o estudo da degradación de biopolímeros utilizados como bioandamios (bioscaffolds) para a rexeneración osea en futuras aplicacións médicas. Para facer o estudo foi preciso a identificación das variables críticas para a o estudo da degradación dos scaffolds. O obxectivo era detectar as variables máis influentes na degradación e medir a súa importancia. Ademais era de gran interese a modelización do chamado camiño de degradación (permitindo a predición do grao de degradación). Para este propósito aplicáronse distintas técnicas estatísticas.

Palabras e frases chave: Biomaterials, Collagen type I, Stem cells, Statistical learning.

1. INTRODUCCIÓN

O uso de modelos estatísticos que permitan estimar como se degradan os materiais ó longo do tempo, en función de variables como o pH ou a temperatura, son de gran interese para futuras aplicacións en medicina rexenerativa, dentro do campo denominado enxeñaría tisular. Concretamente neste traballo presentaremos un par de casos de estudo que foron o centro da tese de doutoramento defendida recentemente na Universidade da Coruña (Robles, 2019).

A aplicación de varias técnicas estatísticas permitiunos estudar dous tipos de degradación, a saber, aquela debida á hidrólise e a degradación debida ao crecemento celular na súa acción de formar tecidos, e en dous grupos diferentes de materiais, PDLGA e o coláxeno tipo I, (actualmente utilizados coma scaffolds (andamios) en moitas e diversas aplicacións médicas).

2. MODELIZACIÓN ESTATÍSTICA DA DEGRADACIÓN DO PDLGA PARA O SEU USO COMO BIOANDAMIOS

Nesta sección analizamos a modelización dos camiños ou tendencias de degradación, causadas por procesos de hidrólise, correspondentes a diversas formulacións de copolímeros PDLG. O fin desta investigación era avaliar e definir as súas posibles futuras aplicacións biomédicas como andamios ou scaffolds dentais. Para este propósito, a variable resposta foi a perda de masa (mass loss), sendo variables regresoras o pH, a temperatura de transición vítrea (Tg) e a masa de auga absorbida polos distintos biopolímeros.

Os datos foron obtidos a partir de mostras dunha disolución PBS (phosphatebuffered saline solution), definidas por un pH inicial de 7.4 (un dos posibles na cavidade bucal e o correspondente ao sangue), e unha temperatura de 37°C (condicións do corpo humano). A perda de masa foi definida como a variable que caracteriza o nivel de degradación do biopolímero, a variable crítica

que define o seu grao de degradación. A súa relación de dependencia con respecto ao tempo, pH e a formulación química foi modelada mediante o uso de ferramentas pertencentes de Statistical Learning.

Na Figura 1 poden observarse as variables analizadas e as correlación existentes entre a perda de masa do material co tempo, co pH ou ca temperatura.

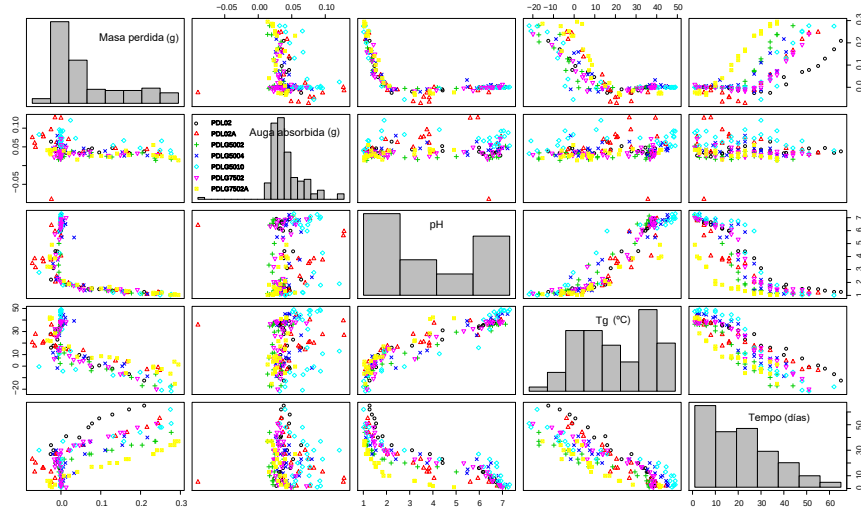


Figura 1: Estudo descriptivo das variables de interese. Poden observarse distintas tendencias e correlacións entre variables.

Nomeadamente, aplicáronse Modelos Aditivos Xeneralizados (GAM) e regresión non linear de efectos mixtos baseada nas funcións loxísticas. Por unha banda, os modelos GAM proporcionan información sobre as variables máis relevantes no grao de degradación e as funcións paramétricas que relacionan perda de masa (resposta do modelo) con respecto ao pH e ao tempo. Por outra banda, os modelos de regresión non linear de efectos mixtos permiten estimar o grao de degradación do material, ademais de definilo cun vector de parámetros con significación física que permite comparar as diversas formulacións de PDLGA. O camiño de degradación para cada polímero PDLGA foi estimado e comparado con respecto aos restantes para axudar a elixir o polímero máis adecuado para cada aplicación médica específica, desenvolvendo criterios de selección.

Como consecuencia do estudo, descubriuse que as diferenzas na perda da masa dos copolímeros PDLGA está fortemente relacionada coa variación do pH en función do tempo, decrecemento debido á presenza crecente de grupos de ácido carboxílico. Isto pode ocorrer naqueles ambientes onde os produtos de degradación fican relativamente estancados e en contacto coa masa non degradada, como é o caso do presente estudo experimental.

Os resultados mostran que o grao de degradación dos polímeros PDLGA, en termos de velocidade de degradación e vida media, aumenta cando existe terminación ácida no polímero, cando se reduce a proporción molar do DL-lactide, cando se reduce a chamada "midpoint viscosity", ou cando non se inclúe o ácido glicólico na formulación.

Na Figura 2 pode verse a aplicación de modelos GAM para estudar a degradación do biopolímero PDLGA. Ademais estudouse o efecto do pH sobre a perda de masa do polímero, é dicir, sobre o seu grao de degradación. Observouse que a relación da perda de masa con respecto ao pH é forte e de tipo asíntotico. Tamén é importante salientar que esta relación non depende do tipo de polímero. Isto é debido ao incremento paulatino da cantidade de ácido carboxílico resultado dos procesos de degradación.

O modelo proposto de mixtura de loxísticas pode expresarse mediante a seguinte ecuación:

$$Mass\ Loss_{ij} = \phi_{1i} + \frac{\phi_{2i} - \phi_{1i}}{\left(1 + \exp\left(\frac{Time_{ij} - \phi_{3i}}{\phi_{4i}}\right)\right)} + \epsilon_{ij}$$

Onde i denota o tipo de polímero usado; ϕ_i sería o vector estimado de parámetros; ϕ_{1i} a masa

inicial da mostra analisada; ϕ_{2i} a masa final do proceso de degradación; ϕ_{3i} : o tempo de máxima degradación; ϕ_{4i} : o punto medio da degradación; sendo ϵ_{ij} o error do modelo.

Como resultado pode afirmarse que cando o pH aumenta, a perda de masa diminúe de forma non lineal (asintótica) e cando a perda de masa aumenta, o pH diminúe independentemente da formulación de PDLGA (ver Figura 2).

Tamén se observa a relación non lineal sigmoide entre perda de masa e tempo. A perda de masa destes biopolímeros usados como andamios dentais depende fortemente do tempo e da formulación de PDLGA (ver Figura 2).

Para maior información sobre este estudo pode consultarse o traballo de Robles-Bykbaev et al. (2018).

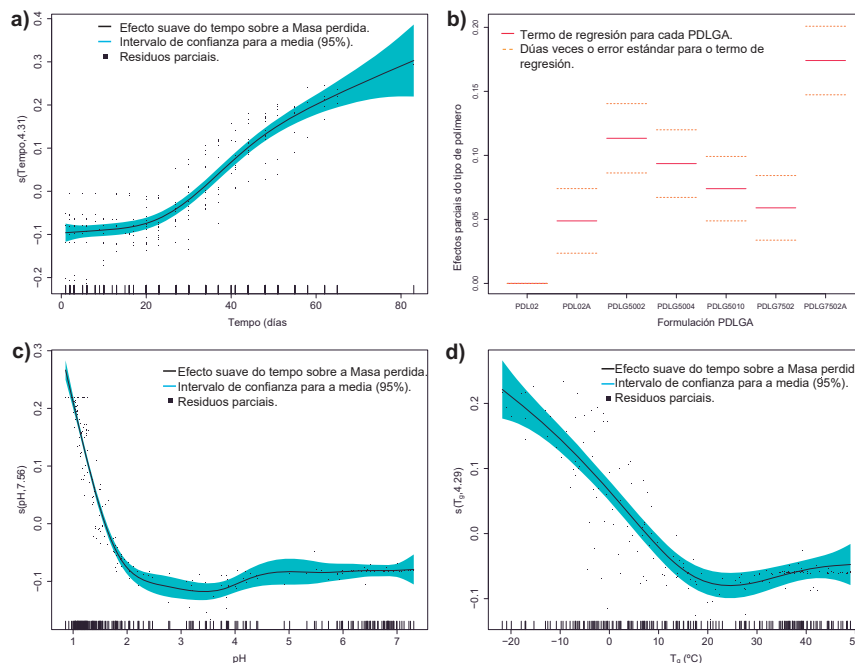


Figura 2: Pode observarse o uso de modelos GAM da perda de masa (a), cos intervalos (b), da pH (c), ou da Tg para todos os polimeros estudados.

3. ESTUDO DA BIODEGRADACIÓN DO COLÁXENO TIPO I USADO PARA A REXENERACIÓN ÓSEA

Nesta sección propónse un método baseado na análise de imaxe para modelar e estimar o grao de degradación do coláxeno tipo I (usado tamén como andamio en rexeneración ósea) debido ao crecemento celular e diferenciación a osteoblastos.

Para a experimentación, que foi levada a cabo en laboratorios do INIBIC, mediuse a masa de coláxeno sometido á acción do crecemento e diferenciación cara osteoblastos a partir de células nai cada 3 días durante un período de 2 meses, "in vitro", baixo condicións que simulan as do corpo humano.

Ademais, aplicáronse técnicas de microscopía óptica para obter información sobre o crecemento celular, a diferenciación celular e a degradación do coláxeno. A primeira contribución consiste na aplicación da clasificación supervisada mediante un algoritmo Random Forest ás características relacionadas coa textura da imaxe (estrutura do tensor e entropía) para estimar as diferentes rexións de interese nunha imaxe obtida por microscopía óptica: a matriz extracelular, o coláxeno, o fondo da imaxe e os núcleos celulares. Acto seguido, se estuda a evolución das rexións correspondentes á matriz extracelular e ó coláxeno a partir da extracción de características da imaxe relacionadas coa progresión do crecemento celular e a degradación do material que serve de andamio (por exemplo, área media dos obxectos da imaxe e a moda do histograma de intensidades).

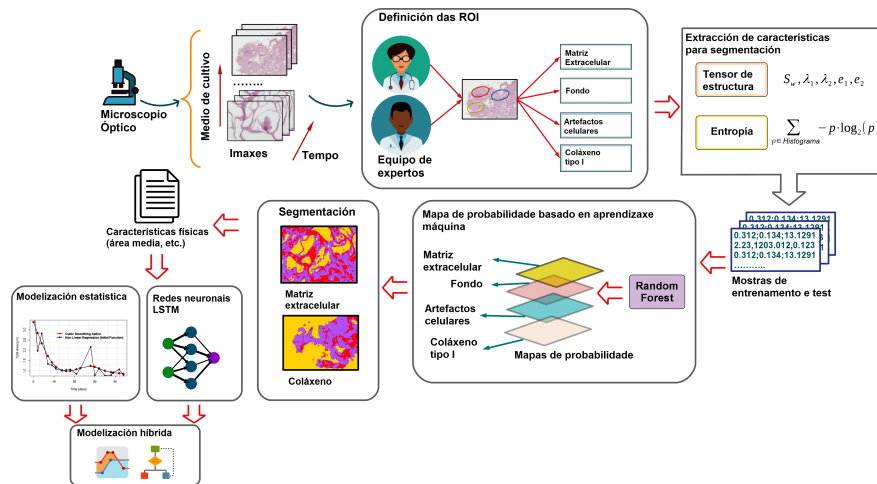


Figura 3: Diagrama usado para o estudo do colágeno mediante segmentación de imáxenes.

Finalmente, estas características críticas modelizáronse estatisticamente con respecto do tempo mediante modelos, tanto non paramétricos como paramétricos, e tanto lineares coma non lineares. En particular, empregouse o modelo de regresión de mestura de funcións loxísticas que proporciona un método que permite separar os procesos de degradación do colágeno debidos á actividade celular ou biolóxica, proporcionando unha estimación da proporción de perda de masa debida a cada causa.

A relación entre o crecemento e diferenciación a osteoblastos a partir de células nai, por unha banda, e a degradación do colágeno, por outra, determinouse e modelouse mediante a análise da circularidade e área dos obxectos de cada imaxe, ademais de a partir do estudo da perda de masa do colágeno.

Este conxunto de técnicas de análise de imaxe, aprendizaxe máquina e aprendizaxe estatística, permítenos caracterizar e parametrizar a biodegradación do colágeno tipo I cando actúa como andamio en rexeneración ósea.

As expresións para o modelo proposto pode verse na seguinte ecuación (neste caso a perda de masa con respecto o tempo):

$$Mass\ Loss_{ij} = \frac{0.3 - 0}{1 + \exp\left(\frac{Time_{ij} - \phi_{3i}}{\phi_{4i}}\right)} + \epsilon_{ij}$$

Para o modelo de degradación do pH con respecto do tempo, a expresión sería a seguinte:

$$pH_{ij} = \phi_{1i} + \frac{\phi_{2i} - \phi_{1i}}{\left(1 + \exp\left(\frac{Time_{ij} - \phi_{3i}}{\phi_{4i}}\right)\right)} + \epsilon_{ij}$$

Nas ecuaciones anteriores ϕ_1 e ϕ_2 denotan os parámetros que poderían tomar valores entre 0 e 0.3 g, respectivamente. O valor inicial para a masa suponse sempre 0 sendo o modelo de perda de tipo asintótico ao final que coincide con 0.3 g (degradación total das mostras de PDLGA).

Os efectos aleatorios estarían indicados por ϕ_1 and ϕ_2 .

Dos resultado de aplicar este modelo resulta que o polímero con una vida media mais alta de degradación hidrolítica coincide co tipo PDL02 (19 días mais que a media, ϕ_3), mentras que o biopolímero cunha menor esperanza de vida sería o PDLG7502A (18 días por debaixo de la media de la poblacin ϕ_3).

Para maior información sobre este estudo onúltese o traballo de Robles-Bykbaev et al. (2019).

4. CONCLUSIÓNS

Neste traballo, propúxose a aplicación da modelización estatística: modelo de regresión tipo GAM e de efectos mixtos para obter información da degradación dos biopolímeros de tipo polilácticos (PDLGA).

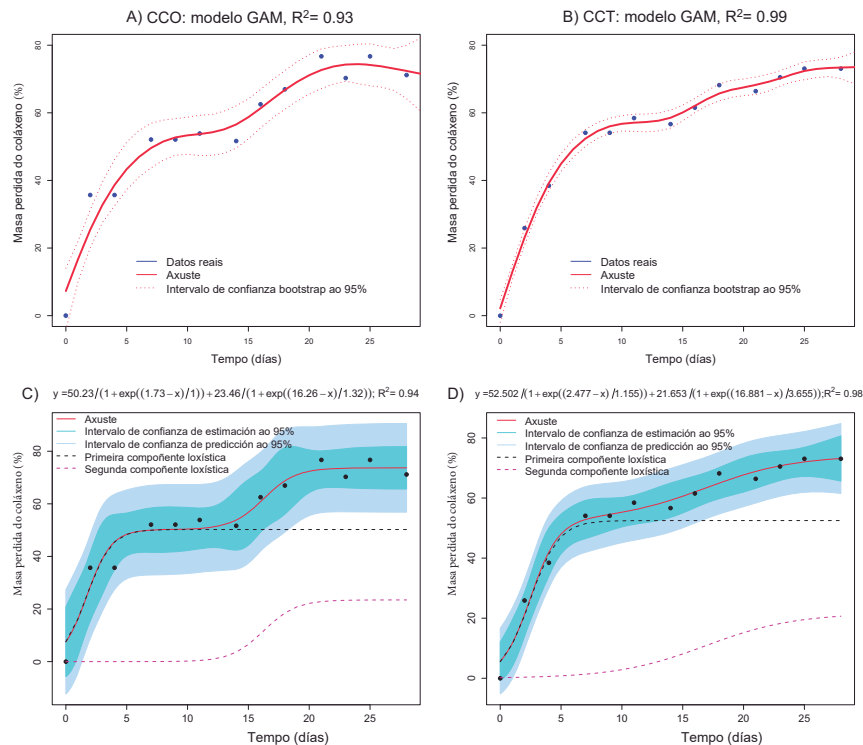


Figura 4: Modelo de mixtura de loxísticas que expresa a perda de masa con respecto do tempo.

Ademais estes modelos permiten comparan a degradación dos diferentes polímeros PDLG. As diferenzas de perda de masa están fortemente relacionadas coa forma en que o pH decae fronte ao tempo.

Tamen se estimaron os parámetros dos modelos para atopar datos sobre a vida media e para caracterizar as tendencias de degradación destes polímeros PDLGA.

Os modelos de regresión de mestura de funcións loxísticas proporcionan un modo de identificar e modelar a degradación debida á actividade biolóxica e, polo tanto, de estimar a correspondente proporción de perda de masa dos biomateriais estudados. Ademais, a metodoloxía proposta na segunda sección pode axudar a estimar o grao de degradación de outros tipos de andamios a partir de imaxes obtidas por microscopía óptica.

5. AGRADECEMENTOS

Os autores agradecen os proxectos nacionais: MTM2014-52876-R e MTM2017-82724-R. Tamén aos proxectos financiados pola Xunta de Galicia: Grupos de Referencia Competitiva ED431C-2016-015 e Centro Singular de Investigación de Galicia ED431G/01. Ademais da axuda da Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) de Ecuador e das axudas predoutorales Inditex-UDC.

REFERENCIAS

- Robles-Bykbaev, Y., Tarrío-Saavedra, J., Naya, N., Díaz-Prado, García-Sabán F.J. and Quintana, S. (2018) Statistical degradation modelling of Poly(D,L-lactide-co-glycolide) copolymers for bioscaffold applications. Plos One. <https://doi.org/10.1371/journal.pone.0204004>.
- Robles-Bykbaev, Y. (2019) Modeling and Analysis of the Degradation of Smart Biomaterials for Dental Use, and the Development and Differentiation of Bone Cells from the Culture of Mesenchymal Cells Inoculated in Dental Scaffolds. Tesis doctoral. Universidade da Coruña.

Robles-Bykbaev, Y., Naya, S., Díaz-Prado, S., Calle-López, D., Robles-Bykbaev, V., Garzón G., Sanjurjo-Rodríguez, C. and Tarrío-Saavedra, J. (2019) An artificial-vision- and statistical-learning-based method for studying the biodegradation of type I collagen scaffolds in bone regeneration systems. PeerJ, DOI 10.7717/peerj.7233 .

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

ANALYSIS OF EXTREME TEMPERATURE EVENTS ON GREENLAND ICE SHEET

María Isabel Borrajo¹, Emma Eastoe² and Amber Leeson³

¹Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.

²Department of Mathematics and Statistics. Lancaster University.

³Lancaster Environmental Centre. Lancaster University.

ABSTRACT

Different climate models have been developed to forecast the future behaviour of climate in polar regions. While these models perform well at reproducing past climates in a broad sense, they have troubles with identifying extreme events. This work focuses on extreme high temperature events on the Greenland ice sheet. An understanding of the location, frequency, duration and magnitude of these events is therefore necessary to understand the ice sheets response to climate change, interpret contemporary measurements of ice sheet volume change and constrain predictions of future ice sheet state. We use advanced statistical techniques for extreme event identification to compile a statistical climatology of extreme temperature events on Greenland based on different regional and global climate models and automatic weather stations data.

Keywords: Extreme value analysis; climate models; Greenland ice sheet; Temperature.

1. INTRODUCTION

Over the next century it is expected that there will be considerable increases in global temperatures. This will have a particular impact in polar regions, where rising air temperature will lead to increased surface snow melt. In turn, this will lead to global sea level rise, to which the Greenland ice-sheet has the potential to be a major contributor. Future ice melt can be predicted, under a range of climate-change scenarios, using global (GCM) and regional (RCM) climate models. Both GCM and RCM are gridded geophysical models, with RCM having a finer scale grid and physics especially developed and tuned for a specific area/climate. Such models are vital in providing an understanding of the local and global consequences of climate change. However the way in which they are constructed focuses exclusively on accurate reproduction of the mean behaviour of the physical processes under consideration with no direct thought given to reproducing the behaviour of the extremes of the physical processes. In the context of ice melt, it is the most extreme temperature events that will lead to the most devastating melt events, and so a lack of consideration of the extremes will be problematic if it turns out that the climate models are not estimating these events well.

2. DATA

In this initial work we analyse data in the period 1979-2017. We have included four GCM: Can-ESM1 (with a spatial resolution of 2.8125 in both, latitude and longitude, which means approximately boxes of side 312km), MIROC5 (with a spatial resolution of 1.40625 in both latitude and longitude, which means approximately 156km); NOR-ESM1-M (with a spatial resolution is 1.875 in latitude, corresponding with approximately 208km, and 2.5 in longitude which is around 278km) and the reanalysis model Era-Interim (with a spatial resolution of 1 in longitude and latitude). Moreover we include two of the most commonly used RCM in Greenland: MAR in its latest version with a spatial resolution of around 15km, and RACMO, with a spatial resolution of around 5.5km.

GC-Net consists of 18 automatic weather stations (AWS) distributed around the ice sheet. We refer the reader to Steffen et al. (1996) for details but briefly summarise here. The first station (Summit) began operation in 1995, with others coming online at various times since then. The AWSs measure a range of meteorological variables, of which the temperature and pressure time series are the most complete. The GC-Net stations each have four temperature sensors (two different instruments mounted at two different heights); here we use data from the type-E thermocouple mounted at position 1 at all sites except NGRIP (North Greenland Ice Core Project), and Saddle during 2010-2016, for which we used data from the type-E thermocouple mounted at position 2. Measurements are taken hourly; we use these data to calculate daily maximum and mean values for compatibility with climate models.

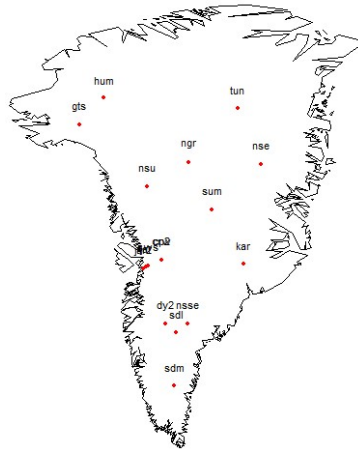


Figure 1: Location of the Gc-Net stations included in the analysis.

3. EXTREME VALUE STATISTICS

There are two main modelling approaches for univariate extreme value analysis: modelling annual maxima (AM) only or taking the peaks-over-threshold (POT) to model all observations above a pre-selected high threshold. The AM approach has the advantage that the extreme events are defined unambiguously, whereas in the POT approach a high threshold must be selected by the user. Conversely, analysing only the annual maxima can potentially lead to a loss of information if it is likely that some years have multiple extreme events, since all but the largest of these multiple events will be discarded. The two approaches are described below.

In this first attempt we use the POT framework, where in our case the threshold is defined as the 90% quantile of the available data. We are studying not only the extremes themselves but their spatial behaviour, hence, for each grid cell in every data set we compute the corresponding threshold. Once the threshold is defined we can classify which space-time positions are extremes and then try to characterise them.

In this context we have performed two different analysis: the first one to compare the models with the AWS data and the second one analysing the spatial behaviour of extreme temperature events in the different climate models.

Compare models with AWS

We have selected some interesting locations of different stations related to their position in the ice sheet and we have extracted the time series of data from the models at those locations. For every time series (stations and models) we have computed mean, maximum and minimum temperature aggregated at different time scales (monthly, seasonal and annual) in the period of study 1979-2017. We provide interesting outputs for a visual comparison. This is an extension of the work developed in Leeson et al (2018).

Spatial extremes in climate models

In this characterisation we are specially interested in three properties of the extremes: number, duration, magnitude and frequency. We compute different descriptive statistics that allow us to plot maps explaining the spatial distribution of extremes.

REFERENCES

Leeson, A., Eastoe, E. and Fettweis, X. (2018). Extreme temperature events on Greenland in observations and the MAR regional climate model. *Cryosphere*, 12, 1091–1102.

Steffen, K., Box, J. E. and Abdalati, W. (1996). Greenland climate network: GC-Net, in: Special report on glaciers, ice sheets and volcanoes, trib. To M. Meier, edited by: Colbeck, S. C., Crrel.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

VARIABILIDADE ONOMÁSTICA VS. VARIABILIDADE LINGÜÍSTICA: UNHA VISIÓN ESPACIAL

M. J. Ginzo-Villamayor¹, R.M. Crujeiras¹, X. Sousa Fernández²

¹ Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela. Tecnoloxías e Análise de Datos Lingüísticos (TECANDALI)

² Departamento de Filoloxía Galega. Universidade de Santiago de Compostela. Tecnoloxías e Análise de Datos Lingüísticos (TECANDALI)

RESUMO

Na caracterización da poboación dunha rexión xeográfica, o estudo dos apelidos preséntase como unha fonte importante de información, que ten sido aproveitado con interesante resultados nas últimas décadas desde distintas disciplinas. O obxectivo deste traballo é comparar os patróns de distribución dos apelidos e variedades dialectais no espazo do dominio lingüístico galego (Galicia e áreas occidentais de Asturias, León e Zamora, véxase Figura 1) con datos lingüísticos da mesma zona.

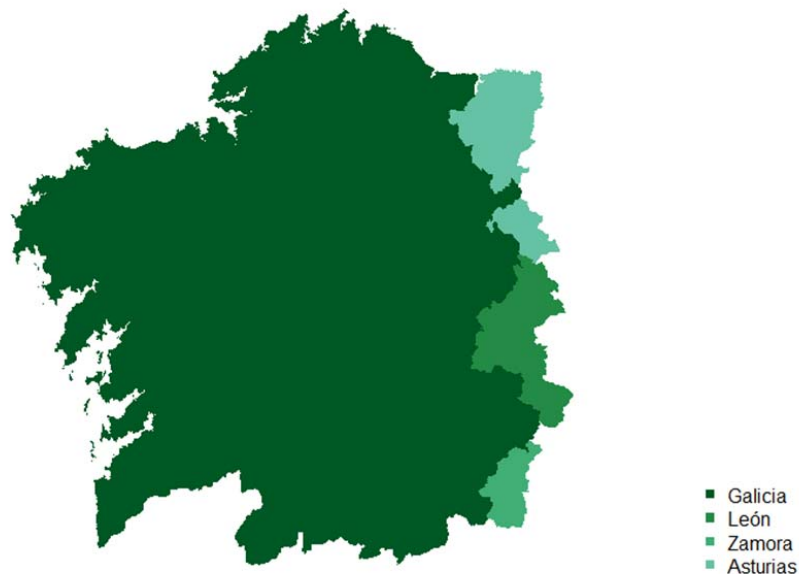


Figura 1: Rexión de estudo formada por 350 concellos.

Nun traballo previo presentado no XI congreso da SGAPEIO (Ginzo-Villamayor, 2013), a proximidade, en termos de similaridade onomástica, entre dous lugares (no noso caso, concellos galegos) puido cuantificarse a través do cómputo dunha distancia, como a distancia de Nei ou a distancia de Lasker.

Resulta tamén de interese analizar a distribución dos apelidos en comparación coa variabilidade lingüística. Por unha banda, a detección de patróns xeográficos similares entre apelidos e dialectos, semella suxerir que os procesos sociais e demográficos de evolución son similares. Por outra banda, que a variabilidade dos apelidos difira da variabilidade

lingüística, pode interpretarse como indicio de que os movementos sociais reflectidos polos dialectos non se corresponden coa historia demográfica da rexión.

As medidas de distancia onomástica entre concellos comparáronse coas distancias lingüísticas calculadas a partir da información recollida no Atlas Lingüístico Galego (ALGa), un proxecto fundacional do Instituto da Lingua Galega (ILG) para describir a riqueza da variación xeolingüística de finais do século (datos de 167 puntos do dominio lingüístico -152 de Galicia e 15 de zonas galegófonas de Asturias, León e Zamora-). Para á análise de apelidos analizáronse 21.213 apelidos diferentes de 2.505.624 de persoas (datos do padrón de 2011, de 350 concellos: 315 galegos, 15 en Asturias, 16 en León e 4 de Zamora) e posteriormente excluíronse aqueles máis comúns que, doutro xeito, darían lugar a unha subestimación dos niveis reais de diversidade.

No estudo empregárase a análise dos compoñentes principais (PCA) que faciliten a identificación gráfica dos posibles patróns de similitude entre os apelidos e os dialectos. Para atopar un patrón de apelidos de Galicia tamén se aplicarán distintas técnicas de agrupación e de identificación de barreiras espaciais, como o algoritmo de Monmonier.

Palabras e frases chave: Medidas de isonimia, distancia Nei, distancia Lasker, PCA, Monmonier.

AGRADECEMENTOS

Os autores deste traballo agradecen ao Instituto Galego de Estatística (IGE), ao Instituto Nacional de Estatística (INE) e a Sociedad Asturiana de Estudios Económicos e Industriales (SADEI) polos datos dos apelidos. Ademais, as investigadoras M^a José Ginzó e Rosa Crujeiras agradecen o apoio do proxecto MTM2016-76969-P do Ministerio de Economía y Competitividad.

REFERENCIAS

- Boattini, A., Blanco-Villegas, M. J., and Pettener, D. (2007). Genetic structure of La Cabrera, Spain, from surnames and migration matrices. *Human Biology*, 79(6), 649–666.
- Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, 19, 233-257.
- Cheshire, J.A. and Longley, P.A. (2012) Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, 26, 309-325.
- Cheshire, J.A., Longley, P.A. and Singleton, A.D. (2010) The surname regions of Great Britain. *Journal of Maps*, 6 401-409.
- De Oliveira, M. Z., Schüler-Faccini, L., Demarchi, D. A., Alfaro, E. L., Dipierri, J. E., Veronez, M. R., Ramallo, V. (2013). So close, so far away: Analysis of surnames in a town of twins (cândido godói, brazil). *Annals of Human Genetics*, 77, 125-136. doi: 10.1111/ahg.12001
- Ginzó-Villamayor, M.J.; Crujeiras, R. and Sousa Fernández, X. (2013). Surname patterns in Galicia. Libro de actas: Libro de Actas do congreso. ISBN: 84-695-8723-4. XI Congreso Galego de Estatística e Investigación de Operacións. A Coruña (España)
- Manni, F., Heeringa, W., and Nerbonne, J. (2006). To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the netherlands. *Literary and Linguistic Computing*, 21(4), 507-527. doi: 10.1093/lilc/fql040
- Mikerezi, I., Shina, E. Scapoli, C., Barbujani, G. Mamolini, E., Sandri, M., Carrieri, A., Rodríguez-Larralde, A. and Barrai, I. (2013) Surnames in Albania: a study of the population of Albania through isonymy. *Annals of Human Genetics*, 77, 232-243.
- Nei, M. (1973) The theory and estimation of genetic distance. *Genetic structure of populations* (ed. N. E. Morton). Hawaii: Hawaii University Press.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

EXTENT OF OCCURRENCE ESTIMATION FOR INVASIVE PLANTS

P. Saavedra-Nieves¹ and A. Rodríguez-Casal¹

¹Universidade de Santiago de Compostela

ABSTRACT

Support estimation deals with the problem of reconstructing the compact support S of an absolutely continuous random vector X from an independent and identically distributed sample $\mathcal{X}_n$ of observations taken in it.

Several proposals for estimating S have been considered in the literature. For instance, if we know that S is convex then the convex hull of the sample is a natural support estimator. But convexity assumption may be too restrictive. For this reason, it is necessary to consider more flexible shape restrictions such as r -convexity for $r > 0$. If we assume that S is r -convex then the r -convex hull of $\mathcal{X}_n$ provides a natural reconstruction of the support. In Rodríguez-Casal (2007), it is proved that if the smoothing parameter r is correctly chosen, the r -convex hull of the sample achieves the same convergence rates (in Hausdorff and distance in measure) as the convex hull. But, in practice, S is unknown and, consequently, the real value of the geometric index r too. Rodríguez-Casal and Saavedra-Nieves (2016) proposed a stochastic algorithm for selecting it from the data under the hypothesis that the sample is uniformly generated. The resulting reconstruction of S is able to achieve the same convergence rates as the convex hull for estimating convex sets.

In this work, we will extend the previous result. We propose a consistent estimator of the optimal value of r such that S is r -convex under the assumption that the sample is generated from a Lipschitz continuous density bounded from below on S . The resulting estimator of S also achieves the same convergence rates as the convex hull for estimating convex sets, but under a much more flexible scenario.

Our support estimator will be considered for reconstructing the *extent of occurrence* (EOO) of invasive plants. Roughly speaking, the International Union for Conservation of Nature defines the EOO of a specie as the convex hull of its geographical locations. We will show that this definition presents some limitations that our estimator is able to overcome.

Key words and phrases: Support estimation, r -convexity, extent of occurrence, invasive plants.

REFERENCES

- Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincaré - Probability and Statistics*, 43, 763–774.
- Rodríguez-Casal, A. and Saavedra-Nieves, P. (2016). A fully data-driven method for estimating the shape of a point cloud. *ESAIM: Probability and Statistics*, 20, 332–348.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

CORRECTION WHEN USING ESTIMATES IN LOGISTIC REGRESSION FOR LONGITUDINAL DATA

Rosa C. S. Oliveira

Matemática, Faculdade de Ciências; CINTESIS, Faculdade de Medicina; Instituto Politécnico
de Bragança, Pólo de Vila Nova de Famalicão, Portugal

ABSTRACT

In this paper, we study nonlinear generalised regression models for analysis of longitudinal data. We indicate a two-stage approach that, in the first stage models a linear model to the longitudinal data to estimate the random effect that characterises the trajectory for each individual and in the second stage uses the estimated intercept and slope as predictors for the covariates used in a probit model. We show that the straight use of the estimates in the probit model produces biased estimates of their outcome. We show how to adapt a Regression Calibration and a Pseudo-Likelihood approaches to this context and compare these approaches with a naive analysis where the estimation error is ignored. Regression Calibration and Pseudo-Likelihood methods seems to be the best choice as they performs well, even for small sample size, if data is not very noisy. It is fair to say, mainly, that the Regression Calibration and Pseudo-Likelihood method perform equally well, nonetheless, Regression Calibration seems to performs a bit better. Nevertheless naive approach has smaller Mean Squared Error, has great absolute bias. Regression Calibration seems to be the best. Our study indicates that correcting for measurement errors instead of falsely assume that errors are not present will produce less bias than ignoring exposure measurement error in the analysis.

Keywords: measurement error, errors-in-variables, general nonlinear model, longitudinal data.

1. INTRODUCTION

The study of risk factors of death is central to health metrics. Biomarkers are objective physical or biologic measures of health conditions. The availability of information about biomarkers has opened the possibility to include biomarkers in population health metrics and in the study of biomarkers as risk factors of death has a longer tradition in health sciences, especially of cardiovascular death or other severe outcomes such as heart attacks. Most evaluations of the diagnostic or prognostic ability of clinical biomarkers focuses on a single measurement and its relation with an outcome of interest. Despite, there are settings where the longitudinal profile of a biomarker may be more informative than an isolated value of cross-sectional observations. One reason sticks from the chances provided by longitudinal data to improve prognostic models for succeeding outcomes for each individual. The motivation problem of this article comes from the SACiUCI study, Póvoa 2002, a prospective observational study aiming to identify demographic and clinical prognostic factors of patients admitted to intensive care with sepsis. In particular, we will revisit the evaluation of the longitudinal profile of CRP as a marker of sepsis resolution. Given that CRP levels increase dramatically in response to an inflammatory process, it is expected that the progressive reduction of these levels would indicate that the systemic inflammatory response that underlies sepsis is becoming under control and thus, associated with a good prognosis for the patient.

We focus on a two-stage approach where first we fit linear model to the longitudinal data to estimate the random effect that characterises the trajectory for each individual and in a second stage, we consider the individual specific intercept and slope of the linear regression as covariates in a model for the binary outcome. Prediction of the outcome variable, e.g. death, is based on

the logistic or probit regression model (e.g.) in the second stage, and as we do not know the true values for the intercepts and slopes but only estimates of these quantities obtained from the linear model, simply substituting the true values by the estimates as covariates in the model for the binary outcomes. will bias the results of the second model. By recognising that the use of the estimates as covariates instead of the true values is identical to a situation where the covariates are measured with error, we can use the same principles of Measurement Error (ME) theory to correct or at least compensate for the bias of the substitution. Tosteson (1998) used a likelihood-based method to fit mixed models in which the covariates contain measurement error. Carroll et al. (2006), on the other hand, used Regression Calibration (RC), a simple method applicable to any regression model and most valuable for estimating parameters in generalised linear models (GLM) with covariate measurement error. However, there is an important difference from the classical error measurement problem. It is usually required that the true values are measured in a subset of the data or, at least, that we have available repeated measurements of the variable measured with error, for identifiability of the error components in the model. Nevertheless, it is well known that naive implementation such as conducting longitudinal data analysis covering measurement error conducts to biased inferences, and taking that into consideration, several methods have been proposed to reduce this bias. Even so, it has not been sufficiently investigated a correction of bias on GLM (a logistic regression model, more precisely) to correct for bias in a linear mixed model adjusted for other covariates. In our setting we do not have the true intercepts or slopes for any of the participants nor multiple measurements of the intercepts and slopes. We will show how to adapt a Regression Calibration (RC) and a Pseudo-Likelihood (PL) approaches to this context and contrast these approaches with a naive analysis where the estimation error is ignored.

3. MODEL SETUP

We start with the formulation of the error measurement. Then we show how to turn our problem into a measurement error problem. Subsequently, we present the covariate model encompassing distributional information of the latent variable and the error model relating the latent variable to the observable measurement (X). Then the response model linking the true to the dependent variable death/survival state (D) in a longitudinal setting are presented. Three methods of estimating the linear measurement model coefficients are discussed: OLS, Pseudo-Likelihood (PL) and RC.

3.1 Data and Basic Model

In the simplest case of across-sectional bivariate regression where the independent variable measured is reported with additive error, the proportional attenuation bias is given asymptotically by $1 - \lambda$, where $\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$.

The quantity λ is often referred to as the reliability of the data. The reliability of a variable in the general case is the slope coefficient from an OLS regression of the correctly measured variable on the mismeasured variable, i.e., the proportion of a change in the observed variable that translates into a change in the true, latent, variable. In this survey we propose a reliability factor when errors are made on the measure/report of a variable in a dichotomous regression scenario.

Estimates which do not account for measurement error are typically biased. Correcting for this bias entails what is often referred to as a bias versus variance tradeoff. What this means is that in most problems, the very nature of correcting for bias is that the resulting corrected estimator will be more variable than the biased estimator. Of course, when an estimator is more variable, the confidence intervals associated with it are longer.

Because $\lambda < 1$ it is clear that while the correction-for-attenuation in β_x reduces its bias to zero there is an increase in variability relative to the variance of the biased estimator β_x . The variability is inflated even further if an estimate $\hat{\lambda}$ is used in place of λ . The price for reduced bias is increased variance, just like in Pseudo-likelihood in which the decrease of bias comes with a price on the variance.

Remember that Mean Squared Error (MSE) is the sum of the variance plus the square of the bias. This is an interesting criterion to use, because uncorrected estimators have more bias but smaller variance than corrected estimators, and the bias versus variance tradeoff is transparent.

We present our model within the framework of counting process. Denote by $i = 1, \dots, n$ individuals, each of whom have a binary disease status D_i . In addition, we have further variables Y_{ij} measured at times $(t_{i1}, \dots, t_{ij})$, with $t_{i1} = 0$ denoting baseline.

We suppose that, unconditionally,

$$\beta_i = (\beta_{0i}, \beta_{1i})^T = \text{Normal}\{\beta = (\beta_0, \beta_1)^T, \Sigma_\beta\}.$$

We link the binary outcome to the continuous measurements via a probit regression, so that

$$\text{pr}(D_i = 1|\beta_i) = \Phi(\alpha_0 + \alpha_1\beta_{0i} + \alpha_2\beta_{1i}). \quad (1)$$

In other words, the binary outcome depends on the true baseline value β_{0i} and the individual slope β_{1i} . While the response variable D_i is observed, the covariates β_i are latent variables, in the sense that we do not observe the true β_i , instead we observe an estimate version of this β_i . There are different estimates we can use and will use OLS.

We do not observe β_i , but when $J > 2$, we can estimate it by $\hat{\beta}_i$, the ordinary least squares estimate of β_i . Write $\mathcal{Y}_i = (Y_{i1}, \dots, Y_{ij})^T$, and

$$V_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{ij} \end{bmatrix}.$$

Notice that $\mathcal{Y}_i = V_i\beta_i + (\epsilon_{i1}, \dots, \epsilon_{ij})^T$. Hence, conditional upon the person,

$$\begin{aligned} \text{cov}(\mathcal{Y}_i | \text{person } i) &= \sigma_\epsilon^2 I_J; \\ \hat{\beta}_i &= \beta_i + (V_i^T V_i)^{-1} V_i^T (\epsilon_{i1}, \dots, \epsilon_{ij})^T. \end{aligned} \quad (2)$$

To understand its properties, recall that the ordinary least squares estimate is

$$\hat{\beta}_i = (V_i^T V_i)^{-1} V_i^T \mathcal{Y}_i.$$

Further, we see that from (2), unconditionally,

$$\begin{aligned} \text{cov}(\beta_i) &= \Sigma_\beta; \\ \text{cov}(\hat{\beta}_i) &= \Sigma_\beta + \sigma_\epsilon^2 (V_i^T V_i)^{-1}; \\ \text{cov}(\hat{\beta}_i, \beta_i) &= \Sigma_\beta. \end{aligned}$$

3.2 Measurement Error Formulation

We now show how to turn our problem into a measurement error problem. We have the model $\text{pr}(D_i = 1|\beta_i) = \Phi(\alpha_0 + \beta_i^T \alpha)$; where $\hat{\beta}_i = \beta_i + U_i$ given $\beta_i = \text{Normal}\{\beta, \Sigma_\beta\}$ and $U_i = \text{Normal}\{0, \Sigma_{ui} = \sigma_\epsilon^2 (V_i^T V_i)^{-1}\}$.

We assume that the measurement error is nondifferential. In addition, unconditionally,

$$E(\beta_i | \hat{\beta}_i) = \mathcal{G}(\hat{\beta}_i, \beta, \Sigma_\beta, \Sigma_{ui}) = \beta + \Sigma_\beta (\Sigma_\beta + \Sigma_{ui})^{-1} (\hat{\beta}_i - \beta); \quad (3)$$

$$\text{cov}(\beta_i | \hat{\beta}_i) = \mathcal{V}(\Sigma_\beta, \Sigma_{ui}) = \Sigma_\beta \{I - (\Sigma_\beta + \Sigma_{ui})^{-1} \Sigma_\beta\}. \quad (4)$$

Finally, since $\int \Phi(a + bx)\phi(x)dx = \Phi\{a/(1 + b^2)^{1/2}\}$, it is easily seen that

$$\text{pr}(D_i | W_i) = \Phi \left[\frac{\alpha_0 + \mathcal{G}^T(\hat{\beta}_i, \beta, \Sigma_\beta, \Sigma_{ui})\alpha}{\{1 + \alpha^T \mathcal{V}(\beta, \Sigma_\beta, \Sigma_{ui})\alpha\}^{1/2}} \right]. \quad (5)$$

3.3 Estimation of $(\sigma_\epsilon^2, \beta, \Sigma_\beta)$

To implement a measurement error analysis, we need to estimate σ_ϵ^2 , β and Σ_β . Define

$$\hat{\sigma}_{\epsilon i}^2 = (J - 2)^{-1} \sum_{j=1}^J (Y_{ij} - V_i \hat{\beta}_i)^2,$$

the Mean Squared Error of the regression for person i . Each $\hat{\sigma}_{\epsilon i}^2$ is unbiased for σ_ϵ^2 , so an unbiased estimate of σ_ϵ^2 is just

$$\hat{\sigma}_\epsilon^2 = n^{-1} \sum_{i=1}^n \hat{\sigma}_{\epsilon i}^2.$$

The unconditional covariance matrix of $\hat{\beta}_i$ is $\Sigma_\beta + \sigma_\epsilon^2 (V_i^T V_i)^{-1}$. Let $\hat{\Omega}$ be the sample covariance of $(\hat{\beta}_1, \dots, \hat{\beta}_n)$. Then since $\hat{\Omega}$ estimates $\Sigma_\beta + \sigma_\epsilon^2 n^{-1} \sum_{i=1}^n (V_i^T V_i)^{-1}$, a consistent estimate of Σ_β is

$$\hat{\Sigma}_\beta = \hat{\Omega} - \hat{\sigma}_\epsilon^2 n^{-1} \sum_{i=1}^n (V_i^T V_i)^{-1}.$$

Finally, off course, an unbiased estimate of β is

$$\hat{\beta} = n^{-1} \sum_{i=1}^n \hat{\beta}_i.$$

3.4 Pseudo-Likelihood Estimation of (α_0, α)

For longitudinal binary data with non repeated and non ignorable missing outcomes over time, a full likelihood approach is very complicated algebraically and MLE can be computationally unfeasible, so, as in this research problem we consider that the outcome is not, always, observed at all times we use PLs to estimate the covariate effects on the marginal probabilities of the outcomes. This method in the probit case is difficult to compute, performs poorly in moderate sample sizes and is non-robust to misspecification on the conditional density $f_{X|W}$. The PL requires specification of the distribution for the data at all pairs of times on the same subject, but makes no assumptions about the joint distribution of the data at three or more times on the same subject, so the method can be considered semi-parametric. To formulate a full likelihood for non-ignorable non-monotone missing outcomes over time, one must specify a model for the repeated binary outcomes of interest, and a model for the missingness mechanism. To estimate the parameters, a full likelihood approach has many nuisance parameters, is complicated algebraically, and maximum likelihood estimation can be computationally prohibitive, especially when the number of times is large. We propose a 'Pseudo-Likelihood' to estimate the covariate effects on the marginal probabilities of the outcomes, in addition to the association parameters and missingness parameters. The PL requires only partial specification of the distribution of observations and missingness indicators at pairs of times, and can be much less computationally prohibitive than maximum likelihood.

3.4 Pseudo-Likelihood Estimation of (α_0, α) - General Case

Define

$$\mathcal{K}(\alpha_0, \alpha, \hat{\beta}_i, \beta, \Sigma_\beta, \Sigma_{ui}) = \frac{\alpha_0 + \mathcal{G}^T(\hat{\beta}_i, \beta, \Sigma_\beta, \Sigma_{ui}) \alpha}{\{1 + \alpha^T \mathcal{V}(\beta, \Sigma_\beta, \Sigma_{ui}) \alpha\}^{1/2}} = \mathcal{K}_i(\cdot)$$

Given the estimates formed, we will estimate (α_0, α) by maximizing the pseudo-loglikelihood

$$\sum_{i=1}^n Y_i \log[\Phi\{\mathcal{K}(\alpha_0, \alpha, \hat{\beta}_i, \hat{\beta}, \hat{\Sigma}_\beta, \hat{\Sigma}_{ui})\}] + \sum_{i=1}^n (1 - Y_i) \log[1 - \Phi\{\mathcal{K}(\alpha_0, \alpha, \hat{\beta}_i, \hat{\beta}, \hat{\Sigma}_\beta, \hat{\Sigma}_{ui})\}]$$

in (α_0, α) . This can be performed either by an optimizer, or by Fisher's method of scoring, illustrated as follows.

The Fisher Scoring algorithm is an iterative maximum likelihood procedure and can be implemented fitting iteratively reweighted least square. Is a similar method to Newton-Raphson but modified to overcome the convergence problem of the Newton-Raphson method. Both algorithms

give the same parameter estimates; however, the estimated covariance matrix of the parameter estimators can differ slightly. This is due to the fact that Fisher scoring is based on the expected information matrix while the Newton-Raphson method is based on the observed information matrix. In the case of a binary logit/probit model, the observed and expected information matrices are identical, resulting in identical estimated covariance matrices for both algorithms.

Define Q_i

$$\begin{pmatrix} \partial \mathcal{K}_i(\cdot)/\partial \alpha_0 \\ \partial \mathcal{K}_i(\cdot)/\partial \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \{1 + \boldsymbol{\alpha}^T \mathcal{V}(\boldsymbol{\beta}, \Sigma_{\beta}, \Sigma_{ui}) \boldsymbol{\alpha}\}^{-1/2} \\ \mathcal{G}_i(\cdot)/\{1 + \boldsymbol{\alpha}^T \mathcal{V}_i \boldsymbol{\alpha}\}^{1/2} - \{\alpha_0 + \mathcal{G}_i^T(\cdot) \boldsymbol{\alpha}\} \{\boldsymbol{\alpha} \mathcal{V}_i\}^T / \{1 + \boldsymbol{\alpha}^T \mathcal{V}_i \boldsymbol{\alpha}\}^{3/2} \end{pmatrix}$$

It is readily seen that the Fisher Score is

$$\mathcal{S}_i(\cdot) = \sum_{i=1}^n \frac{\varphi\{\mathcal{K}_i(\cdot)\} Q_i}{\Phi\{\mathcal{K}_i(\cdot)\} [1 - \Phi\{\mathcal{K}_i(\cdot)\}]} \times [D_i - \Phi\{\mathcal{K}_i(\cdot)\}],$$

while the Fisher Hessian will be:

$$\mathcal{H}_i(\cdot) = - \sum_{i=1}^n \frac{\varphi^2\{\mathcal{K}_i(\cdot)\} Q_i Q_i^T}{\Phi\{\mathcal{K}_i(\cdot)\} [1 - \Phi\{\mathcal{K}_i(\cdot)\}]}.$$

The Fisher scoring update is

$$(\alpha_{0\text{new}}, \boldsymbol{\alpha}_{\text{new}}^T)^T = (\alpha_{0\text{old}}, \boldsymbol{\alpha}_{\text{old}}^T)^T - \{\mathcal{H}(\cdot)\}^{-1} \mathcal{S}(\cdot),$$

where $\mathcal{H}(\cdot)$ and $\mathcal{S}(\cdot)$ are evaluated at $(\alpha_{0\text{old}}, \boldsymbol{\alpha}_{\text{old}}^T)^T$. The covariance matrix of $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}^T)^T$ can be estimated by the bootstrap, thus taking into account the estimation of the parameters. It is possible to use estimating equation theory as well to derived estimates of the asymptotic covariance matrix.

Finally, the measurement error model has to be specified. It relates the CRP measurements to their evolution and baseline value, i.e., the surrogate covariate W_{it} to the latent variable X_{it} . Lastly we have to stipulate the model that relates this CRP evolution to the prediction of disease.

4. RESULTS - ANALYSIS OF SACiUCI DATA

Póvoa et al.(2002) described a prospective, cohort, multi-centered study, conducted over one year (1 December, 2004 to 30 November, 2005) of biomarkers as prognostic factors in SEPSIS. Data of community-acquired sepsis study was collected in seventeen intensive care units (ICU) in Portugal. All adult patients (age > 18 years) consecutively admitted in the participating ICUs were enrolled and screened for CAS. Patients were then followed up until death or hospital discharge. Sepsis and sepsis-related conditions were defined according to the criteria proposed by the American College of Chest Physicians/Society of Critical Care Medicine (Bone, 1992).

Using data from the Portuguese Sepsis Adquirida na Comunidade e internada em Unidade de Cuidados Intensivos (SACiUCI) study (Póvoa, 2002) we investigated the relationship between serial CRP measurements after prescription of antibiotics and the clinical course of Community-Acquired Sepsis (CAS) admitted in Intensive Care Units (ICU). To build the relation, we consider the five measurements of CRP within the GLM model framework and the mortality (D). Only patients with available measurements at Day 1 till Day 5 were considered for this analysis. Patients were followed-up during the first five ICU days, day of ICU discharge or death and hospital outcome. CRP-ratio was calculated in relation to Day 1 CRP concentration. Patients were classified according to the pattern of CRP-ratio response to antibiotics: fast response if Day 5 CRP-ratio was < 0.4, slow response if Day 5 CRP-ratio was between 0.4 and 0.8, and no response if Day 5 CRP-ratio was > 0.8. Data collection included demographic data and comorbid diseases. Clinical and laboratory data at the time of hospital admission was recorded. The Simplified Acute Physiology Score (SAPS) II was calculated. Microbiological and clinical infectious data were reported. C-reactive protein, body temperature, White Blood Cell count (WBC) and the Sequential Organ Failure Assessment (SOFA) score were evaluated during the first five days of ICU stay. Blood samples were obtained from an arterial line at ICU admission and subsequently every morning. For purposes of the time

Táboa 1: Regression Estimates from SACiUCI data.

Estimator	Estimate	SE	OR	$OR_{95\%} CI$	
				LL	UL
CRP init (α_1)					
<i>Naive</i>	.099	.253	1.576	(.960, 2.588)	
<i>Regression Calibration</i>	.105	.270	1.554	(.916, 2.636)	
<i>Pseudo – Likelihood</i>	.209	.074	1.13	(1.04 , 1.23)	
CRP change (α_2)					
<i>Naive</i>	.292	.747	13.115	(3.031, 56.755)	
<i>Regression Calibration</i>	.330	.845	18.187	(3.474, 95.197)	
<i>Pseudo – Likelihood</i>	1.230	.23	2.06	(1.580, 2.69)	

dependent analysis Day 1 (D1) was defined as the day of ICU admission. The following days were successively named as D2 up to D5. The withdrawal of the inflammatory stimulus results in a sharp decrease of the CRP serum concentration in a way similar to a first-order elimination kinetics and, in that sense, relative CRP variations are more informative than absolute changes. As a result we fitted a linear model for the CRP for each patient, i , and assessed the CRP baseline value (intercept) and change described by the slope in the model. Subsequently, we considered a prognostic model using the change of CRP, $P(D_i = 1) = g(\beta_0 + \hat{\beta}X)$ using $probit^{-1}$ as the link function.

The uncorrected analysis of these data were published by Póvoa et al. (Póvoa, 2011). Table 1 compares standard methods to the enhancements to RC and Pseudo-Likelihood discussed in this paper. The results from the uncorrected regression of Sepsis on X , the CRP baseline value and change, taken to be perfectly measured with a binary indicator for disease use the study data in an ordinary logistic regression model. The point estimate $\hat{\beta}$, its estimated standard error ($SE(\hat{\beta})$), the odds ratio and its 95% confidence interval are given. The bias-corrected point estimates were all three-fold smaller than their uncorrected counterpart. Note that the corrected CRP change estimate is approximately 10% smaller than the usual estimate. We are convinced that this is optimistic because the corrected SE estimator has been found to be smaller, approximately 10%, than the standard counterpart.

In this example, the augmented coefficient for CRP change index after correction can be interpreted as meaning that the confounding effects of CRP change in Sepsis prediction are only partially accounted for if one does not correct for measurement error.

5. CONCLUSIONS

The present article presents an efficient method for repeated estimation in a nonlinear model scenario. In contrast to previous works, we studied a robust method for consistent estimation in logistic/probit regression models when *measurement errors* exist in the predictors and the response is binary. Contrary to previous studies the *measurement error* distribution does not need to be known neither replicate data need to be available. The methods discussed in this paper assume non-differential *measurement error*, required for valid application of Regression Calibration method. Though, this assumption is usually fulfilled in prospective studies.

Our simulations confirm that even when the original model could be nicely estimated, the presence of the *measurement error* could severely bias the estimates if the surrogate are used in the estimation. It was shown that the performance of the various methods follows the patterns suggested by the analytic results.

Naive approach is, without doubt, the worse estimator, concerning MSE and bias.

Our simulations show that compared to OLS, RC and PL offers little or no advantage when sample sizes are small, but performs best when samples are reasonably large and especially when the measurement error or the effects are not small.

The OLS method performed poorly, with OLS overestimating the true values consistently except concerning the covariate measured without error, in which case, underestimate is performed.

Compared to the naive OLS approach, Regression Calibration grossly attenuates the estimated error in prediction, i.e, moreover the magnitudes of these estimates are relatively close to the true coefficients, especially respecting the intercept. Concerning greater λ values, i.e., $\lambda > 0.90$ estimates are very good. However, one would not have known this before performing this measurement error correction.

RC outperforms the PL estimator, due in part to the large number of nuisance parameters, which must be managed for the PL estimator. Both the RC and the PL estimators may have non-negligible bias for large sample sizes.

Even though PL approach is more sophisticated it seems to perform worse, especially for small samples, i.e., under rare disease conditions, as it varies profusely and, nevertheless bias "absolute value" decreases, it is unstable for small sample size, probably because when calculating PL we take into account many estimates which makes the coefficients estimates unstable.

The naive OLS has larger bias but smaller SE than Regression Calibration and Pseudo-Likelihood.

Generally, it is fair to say that the Regression Calibration and Pseudo-Likelihood method perform equally well, nonetheless, Regression Calibration seems to perform a bit better. In fact, predictably requires less than 10 Fisher-Scoring iterations to maximize the likelihood. Nevertheless, in the case of one variable, curiously, it is more efficient to use the OLS. We conclude that correcting for measurement errors instead of falsely assume that errors are not present will produce less bias than ignoring exposure measurement error in the analysis.

Our work was novel therein to the current time methods concerning measurement error invoked assumptions difficult to verify. Our method only rests on the assumption of non-differential measurement error and normality of the distribution. Regression Calibration method has gained some popularity because it's simplicity, and nevertheless, is an approximate we showed it performs well when compared to likelihood methods, an interesting alternative that many researchers put apart due the difficulty of likelihood calculation in some designs. There continue to exist a huge number of unanswered questions about correction for covariate measurement error in nonlinear models when applied to longitudinal data or repeated measures. For instance, when there is more than one variable or when the number of covariates is greater than the number of measurements and for so, deserve further study. The first question is nowadays one of our research aims.

REFERENCES

- Bone R.C.; Balk R.A.; Cerra F.B.; Dellinger R.P.; Fein A.M.; Knaus W.A.; Schein R.M. and Sibbald W.J. (1992) Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee, American College of Chest Physicians/Society of Critical Care Medicine, 101, 1644-1655.
- Carroll, R.J.; Ruppert, D.; Stefanski, L.A. and Crainiceanu, C.M. (2006) Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall.
- Póvoa, P. (2002) C-reactive protein: a valuable marker of sepsis. Intensive Care Medicine, 28 (3), 235–243.
- Póvoa, P.; Teixeira-Pinto A. M.; Carneiro, A. H. and Portuguese Community-Acquired Sepsis Study Group SACiUCI (2011). C-reactive protein, an early marker of community-acquired sepsis resolution: a multi-center prospective observational study. Critical Care, 15 (4).
- Rosner, B.; Spiegelman, D. and Willett, W. C. (1990) Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. American Journal of Epidemiology, 132, 734–745.
- Tosteson, T. D.; Buonaccorsi, J. P. and Demidenko, E. (1998) Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. Statistics in Medicine, 17 (17), 1959–1971.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

R PACKAGES TO MISSING LONGITUDINAL DATA

Susana R. G. Martins¹

¹Institute Polytechnic of Viana do Castelo and ISAG-European Business School

ABSTRACT

Statistical studies that extend over time have gained more importance. Thus, longitudinal studies have gained more protagonism. One of the problems with this type of study is the lack of data at certain times. Thus, the study of missing data in longitudinal studies is a theme of special importance. In this text will be made a short compendium of the methodologies available in R for the analysis of missing data in longitudinal studies.

Key words and phrases: R packages, longitudinal data, missing values.

REFERENCES

Cran, <https://cran.r-project.org/>, September 2019.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

NONPARAMETRIC ESTIMATION OF THE CURE RATE WHEN CURE STATUS IS PARTIALLY KNOWN

Wende Clarence Safari¹, Ignacio López-de-Ullibarri² and María Amalia Jácome³

¹ Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, A Coruña, Spain. Email: wende.safari@udc.es

² Department of Mathematics, Escuela Universitaria Politécnica, Universidade da Coruña, Ferrol, Spain. Email: ignacio.lopezdeullibbarri@udc.es

³ Department of Mathematics, Faculty of Science, Universidade da Coruña, A Coruña, Spain. Email: maria.amalia.jacome@udc.es

ABSTRACT

Survival cure models are used to analyse time-to-event data in which the event of interest will never occur to some subjects even if the follow-up is indefinitely long. Standard cure models typically assume that cure status is not observed, hence, implicitly model it as a latent variable. However, there are situations for which some individuals are observed to be cured or insusceptible. For instance, their survival times exceeds a known threshold or given certain prior information e.g. the result of a diagnostic test. Here we propose a nonparametric estimator of the cure rate when the information of the cure status is available. The proposed estimator is shown to be strongly consistent and asymptotically normally distributed. A bootstrap procedure for bandwidth selection is described. Finally, the proposed method is illustrated with an application to a real dataset and a numerical example.

Keywords: Bandwidth; Beran estimator; Bootstrap; Censoring; Cure models; Kernel estimator; Local weights.

1. INTRODUCTION

When there is scientific reasoning or empirical evidence that a proportion of subjects will not experience the event of interest, it may be appropriate to explicitly allow for the cure fraction to be estimated and directly modeled. For example, due to advances in cancer treatment there is an increasing fraction of patients who do not present the failure event even after an extended follow-up; such observations correspond to cured patients or long-term survivors. The use of standard survival analysis for such data is inappropriate because it does not allow to estimate the cure fraction.

Let Y be the time to the event and X a covariate that might have an effect on the variable Y . We set $Y = \infty$ if the subject is cured. The probability of being cured when the covariate $X = x$ is $1 - p(x) = P(Y = \infty \mid X = x)$, and the conditional survival function of the uncured individuals given $X = x$, also called latency, is $S_0(t \mid x) = P(Y > t \mid Y < \infty, X = x)$. The mixture cure model, introduced by Boag (1949) and extensively studied in recent years (see Maller and Zhou (1996) for a detailed review) assumes that the population is split into two groups, the cured and the susceptible to the event. As a consequence, the conditional survival function of Y , $S(t \mid x) = P(Y > t \mid X = x)$, can be written as:

$$S(t \mid x) = 1 - p(x) + p(x)S_0(t \mid x).$$

The estimation of the latency and the probability of being cured, conditional on $X = x$, is usually performed by assuming parametric or semiparametric models.

When analyzing time-to-event data, it is often the case that the lifetime of some subjects is censored. That is, follow-up ends before the event occurs. Under right random censoring, nonparametric estimators for the cure rate (Xu and Peng, 2014; López-Cheda et al. 2017a) and the latency function (López-Cheda et al. 2017b).

Many of the previous works on cure models analyzed the cure status as a latent variable, since it is unknown under censoring. However, if information of the cure status is available, then the observations with known cure status can be explicitly modeled. One typical example is the case when individuals are assumed to be cured if their survival time is larger than a given threshold (e.g., 5 years when considering recurrence in some types of cancer).

Only few authors have explored cure models which allow cure to be partially known. Laska and Meisner (1992) proposed nonparametric estimators of the cure rate and latency when cure status is known if an individual does not develop an event of interest within a pre-specified time period. However, they did not incorporate covariates. Betensky and David (2001) proposed nonparametric estimation methods in the presence of random cure times and competing risk. They defined hospital discharge as the cure event and death as a competing event, but likewise did not incorporate covariates. Nieto-Barajas and Yin (2008) proposed a Bayesian semi-parametric approach for estimating survival probability with a cure fraction in the presence of covariates. Wu et al. (2014) extended the Cox proportional hazards cure model in the possibility of knowing the cure status for an individual based on a diagnostic test with a given sensitivity and specificity. More recently, Bernhardt (2016) proposed a cure rate model which accommodates different censoring distributions for the cured and uncured groups, also allows for some individuals to be observed as cured when their survival time exceeds a known threshold. The model is developed for a general case where any parametric distribution can be used to model $S_0(t | x)$ and provide an expectation-maximization (EM) algorithm for estimating the model parameters. He also showed in simulations that ignoring a cure threshold when one exists can lead to biased estimates.

None of the aforementioned works considered a completely nonparametric cure model in the presence of known cure fraction and covariates. This motivates us to introduce a nonparametric cure rate estimator in the presence of known cure. In Section 2, notation when cure status is available is defined, the cure rate model is presented and some properties are given. For the choice of the bandwidth, we recommend the use of a selector based on the bootstrap. Additionally, in Section 3 we provide the asymptotic results of our proposed estimator. In Section 4 we study the efficiency of our estimator using a simulation study and we compare our estimator with a nonparametric estimator that ignores the cure status. The estimator is applied to a sarcoma dataset in Section 5.

2. METHODOLOGY

Let Y be the time to event of interest and C a right censoring random time with conditional distribution function $G(t | x)$. Furthermore, it is assumed that C is independent of Y given X . Let $T = \min(Y, C)$ be the observed time with distribution function $H(t | x)$, $\delta = \mathbf{1}(Y \leq C)$ the uncensoring indicator, and $\nu = \mathbf{1}(Y = \infty)$ the cure indicator.

Suppose that extra information about the cure status is provided. Let ξ be a binary variable indicating whether the cure status is known ($\xi = 1$) or not ($\xi = 0$). We assume that cure status is independent of Y , C and X . The observations are $\{(X_i, T_i, \delta_i, \xi_i, \xi_i \nu_i), i = 1, \dots, n\}$, and can be classified into three groups: (a) the individual is observed to have experienced the event and therefore known to be uncured ($X_i, T_i = Y_i, \delta_i = 1, \xi_i = 1, \xi_i \nu_i = 0$); (b) the lifetime is censored and the cure status is missing ($X_i, T_i = C_i, \delta_i = 0, \xi_i = 0, \xi_i \nu_i = 0$); and (c) the lifetime is censored and the individual is known to be cured ($X_i, T_i = C_i, \delta_i = 0, \xi_i = 1, \xi_i \nu_i = 1$).

Given the observations, the proposed estimator of $1 - p(x) = P(Y = \infty | X = x)$ is,

$$1 - \hat{p}_h^c(x) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\delta_{[i]} B_{h[i]}(x) + \sum_{j=i+1}^n B_{h[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 0) + B_h^c(x)} \right)$$

where $B_h^c(x) = \sum_{j=1}^n B_{h[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 1)$ is the sum of the weights of all the individuals known

to be cured,

$$B_{h[i]}(x) = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^n K_h(x - X_{[j]})} \quad (1)$$

are the Nadaraya-Watson weights with $K_h(u) = h^{-1}K(u/h)$ a rescaled kernel with bandwidth h . Finally, $\delta_{[i]}$, $X_{[i]}$, $\xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} < T_{(2)} < \dots < T_{(n)}$.

It can be proved that $1 - \hat{p}_h^c(x)$ is the nonparametric local maximum likelihood estimator of the cure rate. The proposed estimator of the cure rate has the following properties:

- (a) If there is no known cure status, it reduces to the nonparametric cure rate estimator proposed by Xu and Peng (2014) and López-Cheda et al. (2017):

$$1 - \hat{p}_h(x) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\sum_{j=i}^n B_{h[j]}(x)} \right).$$

- (b) If there is no censoring, then $1 - \hat{p}_h^c(x)$ reduces to $B_h^c(x)$.

- (c) In an unconditional setting it reduces to

$$1 - \hat{p}_n^c = \prod_{i=1}^n \left(1 - \frac{\mathbf{1}(\delta_{[i]} = 1)}{1 + \sum_{j=i+1}^n \mathbf{1}(\xi_{[j]}\nu_{[j]} = 0) + m} \right)$$

where m is the number of subjects known to be cured, and $n = n_1 + m$. In the particular case, where an individual is known to be cured only if the lifetime is greater than a known fixed time, say d , then the estimator reduces to the nonparametric generalized maximum likelihood estimator of the cure rate as proposed by Laska and Meisner (1992):

$$\hat{p}_n^{LM} = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]}}{\delta_{[i]} + (n_1 - i) + m} \right),$$

where $n = n_1 + m$ with n_1 the number of observed times smaller than d .

- (d) If there exists a common specific known cure threshold, it reduces to the nonparametric cure rate estimator by Xu and Peng (2014) and López-Cheda et al. (2017a).

3. ASYMPTOTIC RESULTS

We present some asymptotic properties of the estimator of the cure rate $1 - p(x)$, showing its consistency and asymptotic normality. Before we state our results, we introduce some notation and a list of technical assumptions. Consider the functions

$$\begin{aligned} H(t | x) &= P(T \leq t | X = x) \\ H^1(t | x) &= P(T \leq t, \delta = 1 | X = x) \end{aligned} \quad (2)$$

$$\begin{aligned} H^{11}(t | x) &= P(T \leq t, \xi = 1, \nu = 1 | X = x) \\ J(t^- | x) &= 1 - H(t^- | x) + H^{11}(t | x) \end{aligned} \quad (3)$$

To prove the asymptotic properties, the following assumptions will be assumed throughout:

- (A1) X , Y and C are absolutely continuous random variables.
(A2) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of the density function of X , $m(x)$, and $I_\delta = [x_1 - \delta, x_2 + \delta]$ for some $\delta > 0$ such that

$$0 < \gamma = \inf\{m(x) : x \in I_\delta\} < \sup\{m(x) : x \in I_\delta\} = \Gamma < \infty$$

and $0 < \delta\Gamma < 1$. For all $x \in I$ the random variable Y is independent of C given $X = x$.

(b) There exist $a, b \in \mathbb{R}$ satisfying

$$\inf \{J(t|x) : x \in I_\delta\} \geq \theta > 0 \quad \text{for } (t, x) \in [a, b] \times I_\delta.$$

- (A3) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.
- (A4) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.
- (A5) The first derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times I_\delta$.
- (A6) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times I_\delta$.
- (A7) The derivatives first with respect to x and second with respect to t of $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded for $(t, x) \in [a, b] \times I_\delta$.
- (A8) The densities (subdensity) corresponding to the distribution (subdistribution) functions $H(t|x)$ and $H^1(t|x)$ are bounded away from 0 in $[a, b]$.
- (A9) The kernel function K is a symmetrical density vanishing outside $(-1, 1)$ and the total variation of K is less than some $\lambda < \infty$.

Theorem 1 *If the assumptions (A1)–(A9) hold, for $x \in I$ and if the bandwidth $h = h_n$ satisfies $h \rightarrow 0, \ln n/nh \rightarrow 0$ and $nh^5/\ln n = O(1)$, then*

$$\hat{p}_h^c(x) - p(x) = (1 - p(x)) \sum_{i=1}^n \tilde{B}_{hi}(x) \zeta(T_i, \delta_i, \xi_i, \xi_i \nu_i, \infty, x) + R_n(x)$$

with

$$\begin{aligned} \zeta(T_i, \delta_i, \xi_i, \xi_i \nu_i, t, x) &= \frac{\mathbf{1}(T_i \leq t, \delta_i = 1)}{J(T_i^- | x)} \\ &\quad - \int_0^t (\mathbf{1}(T_i \geq v) + \mathbf{1}(T_i < v, \xi_i \nu_i = 1)) \frac{dH^1(v | x)}{J^2(v^- | x)}, \\ \tilde{B}_{hi}(x) &= \frac{1}{m(x)} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \end{aligned} \quad (4)$$

where $m(x)$ is the density function of the covariate X , $H^1(t | x)$ and $J(t^- | x)$ are given in (2) and (3), and the remainder term is

$$\sup_{x \in I} |R_n(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \quad a.s.$$

In Corollary 1 we show that the proposed estimator $1 - \hat{p}_h^c(x)$ is strongly consistent.

Corollary 1 *If the assumptions (A1)–(A9) hold, for $x \in I$ and if the bandwidth $h = h_n$ satisfies $h \rightarrow 0, \ln n/nh \rightarrow 0$ and $nh^5/\ln n = O(1)$, then*

$$\sup_{x \in I} |\hat{p}_h^c(x) - p(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) \quad a.s.$$

Before stating the asymptotic normality of the proposed estimator of the cure rate, the next proposition provides the asymptotic expressions of the bias and variance of $1 - \hat{p}_h^c(x)$.

Proposition 1 *The bias and variance of the nonparametric estimator of the cure rate $1 - \hat{p}_h^c(x)$ are:*

$$\mu_{h,c}(x) = h^2 B_c(x) + O(h^4) \quad \text{and} \quad \sigma_{h,c}^2(x) = \frac{1}{nh} s_c^2(x) + O\left(\frac{h}{n}\right)$$

where

$$B_c(x) = \frac{1}{2} (1 - p(x)) \frac{2\Phi'_c(x, \infty, x) m'(x) + \Phi''_c(x, \infty, x) m(x)}{m(x)} d_K, \quad (5)$$

$$s_c^2(x) = \frac{(1 - p(x))^2}{m(x)} \Phi_1^c(x, \infty, x) c_K, \quad (6)$$

with $m(x)$ the density function of the covariate X ,

$$\begin{aligned} \Phi_c(y, t, x) &= E[\zeta(T, \delta, \xi, \xi\nu, t, x) \mid X = y] \\ \Phi_1^c(y, t, x) &= E[\zeta^2(T, \delta, \xi, \xi\nu, t, x) \mid X = y], \end{aligned}$$

with ζ given in (4), and

$$d_K = \int v^2 K(v) dv \quad \text{and} \quad c_K = \int K^2(v) dv.$$

The function $\Phi_c(y, t, x)$ in the bias term $\mu_{h,c}(x)$ verifies

$$\Phi_c(y, t, x) = \int_0^t \frac{dH^1(v \mid y)}{1 - H(v^- \mid x) + H^{11}(v^- \mid x)} - \int_0^t \frac{1 - H(v^- \mid y)}{(1 - H(v^- \mid x) + H^{11}(v^- \mid x))^2} dH^1(v \mid x),$$

with

$$\begin{aligned} \Phi'_c(x, t, x) &= -\frac{S'(t \mid x)}{S(t \mid x)}, \\ \Phi''_c(x, t, x) &= 2 \int_0^t \frac{G'_c(v^- \mid x)}{1 - G_c(v^- \mid x)} d\left(\frac{S'(v \mid x)}{S(v \mid x)}\right) dv - \frac{S''(t \mid x)}{S(t \mid x)}, \end{aligned}$$

with

$$1 - G_c(t \mid x) = (1 - G(t \mid x)) + \pi_1(t, x)(1 - p(x))G_1(t \mid x)$$

where

$$\begin{aligned} \pi_1(t, x) &= P(\xi = 1 \mid \nu = 1, C \leq t, X = x), \\ G_1(t \mid x) &= P(C \leq t \mid \nu = 1, X = x). \end{aligned}$$

The function and $\Phi_1^c(x, \infty, x)$ in the variance term $\sigma_{h,c}^2(x)$ is

$$\Phi_1^c(x, \infty, x) = \int_0^\infty \frac{dH^1(v \mid x)}{(1 - H(v^- \mid x) + H^{11}(v^- \mid x))^2}.$$

Theorem 2 *Suppose the assumptions given in Theorem 1 are satisfied. Then, for $x \in I$ as $n \rightarrow \infty$, it follows that*

a) If $nh^5 \rightarrow 0$ and $\frac{(\ln n)^3}{nh} \rightarrow 0$, then

$$\sqrt{nh}(\hat{p}_h^c(x) - p(x)) \sim N(0, s_c^2(x))$$

b) If $h = Cn^{-1/5}$ then

$$\sqrt{nh}(\hat{p}_h^c(x) - p(x)) \sim N(C^{5/2}B_c(x), s_c^2(x))$$

with $B_c(x)$ in (5), and $s_c^2(x)$ given in (6).

3.1 BANDWIDTH SELECTION

From Proposition 1, the asymptotic mean squared error (AMSE) of the estimator $\hat{p}_h^c(x)$ is obtained:

$$AMSE(\hat{p}_h^c(x)) = h^4 B_c^2(x) + \frac{1}{nh} s_c^2(x), \quad (7)$$

with $B_c(x)$ and $s_c^2(x)$ are given in (5) and (6) respectively. The bandwidth h minimizing (7) is

$$h_{x,AMSE} = \left(\frac{s_c^2(x)}{4B_c^2(x)} \right)^{1/5} n^{-1/5}. \quad (8)$$

Estimating the unknown functions in (8) is a cumbersome task as it depends on unknown quantities. One possibility to overcome this difficulty is to consider a bootstrap estimation procedure.

The bootstrap bandwidth selector is the bandwidth h minimizing the bootstrap version of the MSE, that can be approximated as follows

$$MSE_x^*(h) \simeq \frac{1}{B} \sum_{b=1}^B (\hat{p}_h^{c,*b}(x) - \hat{p}_{g_x}^c(x))^2, \quad (9)$$

where $\hat{p}_h^{c,*b}(x)$ is the nonparametric estimator of $p(x)$ computed with a bandwidth h and the b th bootstrap resample, and $\hat{p}_{g_x}^c(x)$ is the same estimator computed with a pilot bandwidth g_x and the original sample. Considering a bandwidth search grid $\{h_1, \dots, h_L\}$, the procedure for obtaining the bootstrap bandwidth selector of h for a fixed covariate value x , is as follows:

1. Generate B bootstrap resamples of the form $\{(X_i^{(b)}, T_i^{*(b)}, \delta_i^{*(b)}, \xi_i^{*(b)}, \xi_i^{*(b)} \nu_i^{*(b)}), i = 1, \dots, n\}$, for $b = 1, \dots, B$.
2. For the b th bootstrap resample, compute the nonparametric estimator $\hat{p}_{h_l}^{c,*b}(x)$ with bandwidth $h_l \in \{h_1, \dots, h_L\}$.
3. With the original sample and the pilot bandwidth g_x , compute $\hat{p}_{g_x}^c(x)$.
4. For each bandwidth $h_l \in \{h_1, \dots, h_L\}$, compute the Monte Carlo approximation of $MSE_x^*(h_l)$, given by (9).
5. The bootstrap bandwidth, h_x^* , is the minimizer of $MSE_x^*(h)$ in (9) over the grid of bandwidths $\{h_1, \dots, h_L\}$.

In step 1, we consider the simple weighted bootstrap, without resampling the covariate X , which is equivalent to the simple weighted bootstrap proposed by Li and Datta (2001), and used for bandwidth selection without known cures by López-Cheda et al. (2017a). For a fixed x and $i = 1, \dots, n$, we set $X_i^* = X_i$ and generate $(T_i^*, \delta_i^*, \xi_i^*, \xi_i^* \nu_i^*)$ from the weighted empirical distribution:

$$\hat{F}_{g_x}(u, v, w, y|x) = \sum_{i=1}^n B_{g_x i}(x) \mathbf{1}(T_i \leq u, \delta_i \leq v, \xi_i \leq w, \xi_i \nu_i \leq y)$$

and $B_{g_x i}(x)$ is the Nadaraya-Watson weight in (1) with pilot bandwidth g_x . The resulting bootstrap resample is $\{(X_i, T_i^*, \delta_i^*, \xi_i^*, \xi_i^* \nu_i^*), i = 1, \dots, n\}$.

Regarding the pilot bandwidth g_x needed to generate the bootstrap resample and in step 3, it is typically chosen asymptotically of order larger than h_x . This smoothing pilot bandwidth is required for the bootstrap bias and variance to be asymptotically efficient estimators for the bias and variance terms. The order $n^{-1/9}$ satisfies the conditions for this asymptotically optimal pilot bandwidth in Theorem 1 of Li and Datta (2001), and it coincides with the order obtained by Cao-Abad and González-Manteiga (1993) for the uncensored case. Simulation results in López-Cheda et al. (2017a) showed that the choice of the pilot bandwidth has a small effect on the final bootstrap bandwidth. Consequently, we propose to use the same local pilot bandwidth:

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ is the distance from x to the k th nearest neighbor on the right, $d_k^-(x)$ the distance from x to the k th nearest neighbor on the left, and k a suitable integer depending on the sample size. If there are not at least k neighbors on the right (or left), we use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$) respectively. Our numerical experience shows that a good choice is to consider $k = n/4$. Note that when $n = 100$ the value of the local pilot bandwidth g_x is the mean distance to the 25th nearest neighbour on both the left and right sides.

4. SIMULATION STUDY

We demonstrate the performance of our estimator of cure probability by considering two different simulation scenarios. In Model 1, we assume the cure rate is a logistic function

$$1 - p_1(x) = 1 - \frac{\exp(0.476 + 0.358x)}{1 + \exp(0.476 + 0.358x)},$$

whereas in Model 2 the cure rate is cubic polynomial function

$$1 - p_2(x) = 0.5 - \frac{1}{16000}x^3.$$

The survival function of the uncured subjects is:

$$S_0(t | x) = \begin{cases} \frac{\exp(-\alpha(x)t) - \exp(-\alpha(x)4.605)}{1 - \exp(-\alpha(x)4.605)} & 0 \leq t \leq 4.605 \\ 0 & t > 4.605 \end{cases} \quad \text{where } \alpha(x) = \exp\left(\frac{x+20}{40}\right).$$

The covariate X was generated from a uniform distribution $U[-20, 20]$. The censoring times C were generated from an exponential distribution $\text{Exp}(0.3)$. In this way, the percentage of censoring is 54% with an average cure rate of 47% in Model 1, whereas in Model 2, 61% of the observations are censored and the corresponding cure rate is 50%. In both scenarios, the probability that a cure is observed is $\pi = 80\%$. Data were generated such that the censoring times C and the lifetimes Y are independent conditionally on X . We simulated $m = 1000$ samples of sample size $n = 100$. The number of bootstrap resamples was $B = 1000$.

Figure 1, in Model 1 shows the MSE of the nonparametric estimator of the cure rate when cure is partially known, $1 - \hat{p}_h(x)$, computed in a grid of 100 bandwidths, from $h_1 = 3$ to $h_{100} = 20$, equispaced in a logarithmic scale. The plot also includes the MSE of the cure rate estimator by Xu and Peng (2014), which ignores the information on the known cure status computed with the optimal MSE bandwidth. In Model 2, it can be seen that there are several bandwidths for which the proposed estimator clearly improves or is comparable to the nonparametric estimator ignoring the cure status. A similar performance can be seen in Model 1 for some values of the covariate X but for some range of values between $x = -1$ and $x = 12$, the bias of estimator is higher compared to that of the nonparametric estimator when the observations with known cure status are ignored.

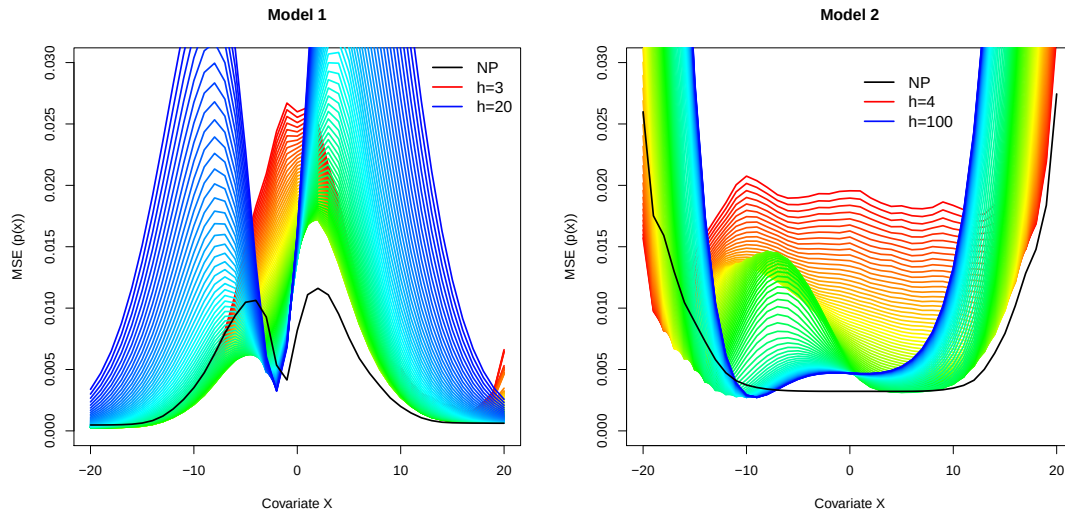


Figura 1: MSE of the proposed nonparametric estimator of the cure rate $1 - \hat{p}_h^c(x)$ in a grid of 100 bandwidths, from $h_1 = 3$ to $h_{100} = 20$ (red) to highest (blue) in Model 1 (left), and from $h_1=4$ (red) to $h_{100}=100$ (blue) in Model 2 (right). The MSE of the cure rate estimator by Xu and Peng (2014) and López-Cheda et al. (2017a) computed with the optimal bandwidth is given as reference (black line).

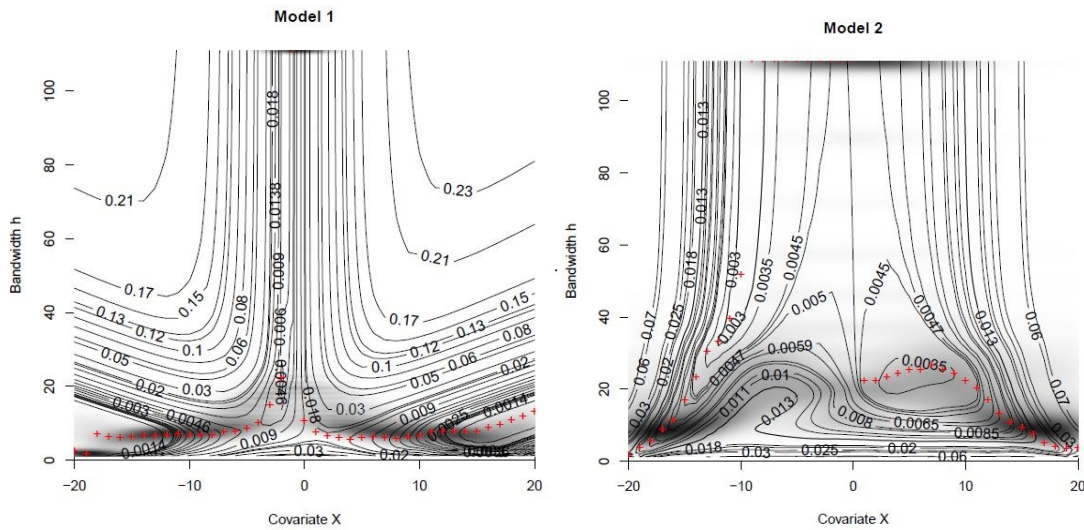


Figura 2: MSE contour plot depending on the bandwidth and the covariate. The density of the bootstrap bandwidth is displayed in the gray-scale, the $h_{x,opt}$ for each covariate is presented in red crosses.

Figure 2 displays the contour plot for the density of the bootstrap bandwidth together with the optimal MSE bandwidth for Model 1 and Model 2. For each value of the covariate X , the theoretical optimal bandwidth, $h_{x,opt}$, is indicated with a red cross. Besides, the lines indicate the MSE as a function of covariate X and the bandwidth h . On the other hand, the density of the bootstrap bandwidths is plotted in a gray area. We see that the density of bootstrap bandwidth, $h_{x,boot}$ performs quite well in estimating the theoretical optimal bandwidth, $h_{x,opt}$. The closer the shaded area to the red crosses the better the bootstrap bandwidth is performing.

5. APPLICATION TO THE SARCOMA DATA

While this estimator can be applied into different research areas, the motivation in this paper comes from a data set related to patients with sarcoma. There were 233 sarcoma patients in the data set. The main outcome was death from sarcoma. A total of 59 (25.2%) patients died from sarcoma, and 174 (74.8%) patients were censored. Within the censored patients, a total of 18 patients were known to be long-term survivors, as they were tumor free for more than five years.

Our estimator is used to estimate the probability of cure conditional on patient's age (ranging from 20 to 90 years). The estimate was compared with that produced by the semiparametric estimator of Bernhardt (2016), which assumes a logistic model for the cure probability and the nonparametric estimator of the cure rate but ignoring the cure status by Xu and Peng (2014) and López-Cheda et al. (2017a).

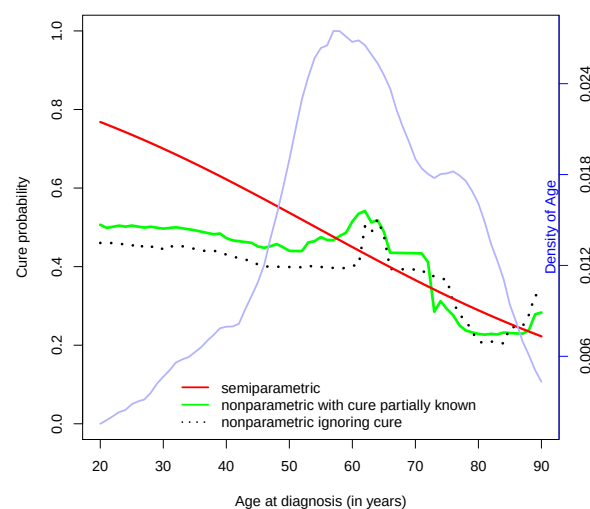


Figura 3: Estimation of the cure probability with the nonparametric estimators computed with bootstrap bandwidths $h_{x,boot}$ (in green and black) and with the semiparametric estimator of Bernhardt (2016) (in red).

Figure 3 shows the cure probability estimates obtained with the proposed nonparametric estimator (in green), the nonparametric estimator ignoring the cure status (in black) and the semiparametric estimator (in red). We see how age effect is relevant to the estimated cure rate.

ACKNOWLEDGEMENT

This work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

REFERENCES

- Bernhardt, P. (2016) A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine* 35, 4607–4623.
- Betensky, R. A., and David A. S. (2001). Nonparametric estimation in a cure model with random cure times. *Biometrics* 57, 282–286.
- Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B (Methodological)* 11, 15–53.
- Cao-Abad, R., and González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics* 2, 379–388.

- Laska, E.M. and Meisner, M.J. (1992) Nonparametric estimation and testing in a cure model. *Biometrics* 48, 1223–1234.
- Li, G., and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Annals of the Institute of Statistical Mathematics* 53, 708–729.
- López-Cheda, A., Cao, R., Jácome, M.A. and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics and Data Analysis* 105, 144–165.
- López-Cheda, A., Jácome, M. A., and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *TEST*, 26(2), 353-376.
- Maller, R. A., and Zhou. *Survival analysis with long-term survivors*. John Wiley & Sons, 1996.
- Nieto-Barajas, L. E., and Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian Journal of Statistics* 35, 540–556.
- Xu, J., and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics* 42, 1–17.
- Wu, Y., Lin, Y., Lu, S. E., Li, C. S., and Shih, W. J. (2014). Extension of a Cox proportional hazards cure model when cure information is partially known. *Biostatistics* 15, 540–554.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

A FUNCTIONAL DIAGNOSTIC TEST BASED ON ROC CURVES ANALYSIS

Graciela Estévez-Pérez ¹ and Philippe Vieu ²

¹Departamento de Matemáticas, Universidade da Coruña

²Institut de Mathématiques, Université Paul Sabatier, Toulouse

ABSTRACT

In this work we investigate diagnostic tests in situations when the observed variables are functional, that is, diagnostic tests that use the functional variables as biomarkers. A procedure based on functional version of ROC analysis is proposed, the main question being linked with a suitable way for ranking the sample of functional data.

Keywords: Test diagnostic; Functional Data; ROC curves.

1. INTRODUCTION

Nowadays, there are many fields, including Economy, Biology and Medicine, that handle functional data, that is, realizations of random elements taking values in an infinite dimensional space. As the analysis of this type of data requires using specific tools to extracting relevant information from them, numerous statistical techniques have been extended to functional context in the two last decades (Ferraty and Vieu, 2006; Horvath and Kokoszka, 2012; Cuevas, 2014; Aneiros et al., 2019). However, the development of diagnostic methods and analysis of ROC curves for functional biomarkers is an exception. In fact, the FDA (functional data analysis) community has started debating recently on this topic but, at our knowledge, only when functional data are involved as covariables (Inacio et al., 2012 and Inacio et al., 2016). The reason for this lack of development for ROC analysis with functional sample is probably linked with the difficulty of establishing some ordering in infinite-dimensional spaces. However, the technological advances are allowing to observe markers with more complex structures and the need for such a functional extension of ROC methodology becomes each day more important in the clinical practice and the medical research. In effect, the functional version of analysis ROC would let design diagnostic test based on functional data. For example, the gene expression levels from a high number of genes could help to discriminate between normal and tumor samples; or the motor speed derived by the Finger Taping Test (FTT), could to discriminate between Parkinson's patients and healthy subjects.

The main purpose of this work is constructing a diagnostic test based on a functional biomarker. It has been possible because of the new general methodology, developed by the authors, for ordering data in infinite dimensional spaces. Specifically, we construct a classification rule, we analyze precisely the concepts of specificity and sensitivity and we discuss how a functional version of ROC curve can help in balancing the trade-off between sensitivity and specificity. Finally, we propose a fully automatic diagnostic testing procedure in which the parameters of the method must be data-driven selected.

In addition, we present some results of a simulation study in R conducted to evaluate the performance of the proposed diagnostic test and to solve the computational issues linked with the choices of the various parameters of the methodology. We also implement some criteria to select the optimal cutoff curve and analyze the accuracy of functional diagnostic test proposed.

2. A FUNCTIONAL DIAGNOSTIC TEST

2.1. The framework and the classification rule

Let $\chi \equiv \chi(\cdot)$ be some functional bio-marker from which we wish to develop testing procedures in order to separate Affected subjects (A) from Non-Affected ones (NA). Let $\chi_1 = \chi/A$ and $\chi_2 = \chi/NA$ be the conditioned functional variables, which are valued in a semi-metric space (E, d) . In addition, we suppose that the probability measures of random variables χ_1 and χ_2 differ by a shift $\Delta \in E$ in the location.

We are in front of a discrimination problem and the data consist in n_1 and n_2 i.i.d. random samples as χ_1 and χ_2

$$\{\chi_{ij}(t); i = 1, 2; j = 1, \dots, n_i; t \in T\}$$

where, without loss of generality, we assume that $T = [0, 1]$ and $n = n_1 + n_2$ denotes the total sample size.

The classification rule: how could we classify an individual as Affected (A) or Non-Affected (NA) from the functional data χ_0 , being χ_0 the evaluation of bio-marker χ on this individual? Our diagnostic method consists in finding a cutoff function or discrimination function $\chi_{c_0} \equiv \chi_{c_0}(t) \in E$, with $t \in T$, which separates Affected (A) and Non-Affected (NA) subjects properly.

Step 1 We select $M_A \equiv M_A(\cdot)$ and $M_{NA} \equiv M_{NA}(\cdot)$ two curves of the functional space E as representatives of each group A and NA , respectively. Once M_A and M_{NA} have been fixed, we consider the totally ordered set $(E_1, \preceq_1)$, where E_1 is the line in E through functions M_A and M_{NA} , that is the family of curves:

$$E_1 = \{\chi_c(t) = cM_A(t) + (1 - c)M_{NA}(t); t \in T, c \in [0, 1]\}$$

and $\preceq_1$ is the total order in E_1 induced by the natural ordering in $\mathbb{R}$.

Step 2 To set a “preorder relation” in the functional space E , we take as pilot ordered subset E_1 , the line in E through M_A and M_{NA} , with a total order relation $\preceq_1$. On the other hand, we consider the projection function $\eta : E \rightarrow E_1$ that assigns each curve in E the nearest element in E_1 when a semi-metric $d(\cdot, \cdot)$ is used. In other words, the projection function is defined by:

$$\eta(x) = \min\{y \in E_1 / d(x, y) = \min\{d(x, y_1); y_1 \in E_1\}\}$$

Note that $\eta(\cdot)$ allows to establish a preorder relation in E ($\preceq$): we say that x_1 precedes to x_2 in E (and write it as $x_1 \preceq x_2$), if and only if $\eta(x_1) \preceq_1 \eta(x_2)$ in E_1 .

Step 3 Predetermined cutoff curve a ζ and assuming that curves with higher values are associated with Affected subjects, **an individual associated with the new functional data χ_0 is diagnosed as Affected (A) if $\chi_0 \not\preceq \zeta$** . In this case we say that the test result is $+$, and we denote it by T^+ .

The classification result depends on the selected curve $\zeta \in E$. To solve this question, we propose to move continuously ζ in $\{\chi_c \in E_1; c \in [0, 1]\}$ and to take the value $c_0 \in [0, 1]$ such as the curve χ_{c_0} lets to discriminate A and NA in the most efficient way.

2.2. Functional ROC curves

Each prefixed curve χ_c provides a dichotomic diagnostic test, which may fail in two different ways:

Test result	True Status	
	Affected	Non-Affected
Positive	True Positive (TP_c)	False Positive (FP_c)
Negative	False Negative (FN_c)	True Negative (TN_c)

In order to have a criterion for data-driven selection of this key parameter c , the sensitivity (*i.e. the true positive fraction* $1 - \beta_c = TPF_c$) and specificity (*i.e. the true negative fraction* $1 - \alpha_c = TNF_c$) can be quantified, as functions of c , as follows:

Specificity

$$1 - \alpha_c = P(T^-/NA) = P_{NA}(\chi \preceq \chi_c) = P_{1,NA}(\eta(\chi) \preceq_1 \eta(\chi_c)),$$

Sensitivity

$$1 - \beta_c = P(T^+/A) = 1 - P_A(\chi \preceq \chi_c) = 1 - P_{1,A}(\eta(\chi) \preceq_1 \eta(\chi_c)),$$

where P denotes the probability law of χ , P_A, P_{NA} the corresponding probability laws of χ_1 and χ_2 , respectively, and $P_{1,A}, P_{1,NA}$ these conditional distributions over E_1 .

How to select the optimal cutoff curve χ_{c_0} ?

The accuracy of the test is defined by the pairs $(1 - \alpha_c, 1 - \beta_c)$ for each cutoff curve χ_c . If the value $c \in [0, 1]$ is moved the sensitivity and the specificity vary inversely, then to selecting the “best” cutoff curve χ_{c_0} , it is needed a balance between sensitivity and specificity.

The **Receiver Operating Characteristic (ROC) curve**, the graph in the unit square representing for each $c \in [0, 1]$ the pair $(\alpha_c, 1 - \beta_c)$, is a global measure of the discriminatory ability of our functional diagnostic test along all cutoff curves χ_c .

$$ROC(c) = (\alpha_c, 1 - \beta_c) = (1 - F_{\eta_2}(\chi_c), 1 - F_{\eta_1}(\chi_c)), \forall c \in [0, 1],$$

where

$$F_{\eta_1}(x) = F_{\eta_A}(x) = P_{1,A}(\eta(\chi) \preceq_1 x)$$

$$F_{\eta_2}(x) = F_{\eta_{NA}}(x) = P_{1,NA}(\eta(\chi) \preceq_1 x)$$

are the distribution functional of $\eta(\chi_1)$ and $\eta(\chi_2)$, for each non-stochastic function $x(\cdot)$.

Estimation of the ROC curve. Based on our data $\{\chi_{ij}(t); i = 1, 2; j = 1, \dots, n_i; t \in T\}$, the most natural procedure for estimating the ROC curve lies in estimating $F_{\eta_A}(\cdot)$ and $F_{\eta_{NA}}(\cdot)$ by their empirical distributions. These estimators are constructed from $\{\eta(\chi_{ij}); i = 1, 2; j = 1, \dots, n_i\}$, i.i.d. observations of $\eta(\chi_1)$ and $\eta(\chi_2)$, respectively. The ROC curve is estimated from the estimations of coefficients α and β :

$$\hat{\alpha}_c = 1 - \hat{F}_{\eta_{NA}}(\chi_c) = 1 - \frac{1}{n_2} \sum_{j=1}^{n_2} I[\eta(\chi_{2j}) \preceq \chi_c],$$

and

$$1 - \hat{\beta}_c = 1 - \hat{F}_{\eta_A}(\chi_c) = 1 - \frac{1}{n_1} \sum_{j=1}^{n_1} I[\eta(\chi_{1j}) \preceq \chi_c].$$

2.4. An optimal data-driven procedure

For selecting **the optimal cutoff curve χ_{c_0} or discrimination curve**, we propose to take χ_{c_0} such that the pair of values $(1 - \alpha_{c_0}, 1 - \beta_{c_0})$ is as close as possible to the pair of perfect classification given by $(1, 1)$. This is equivalent to searching **the point on the ROC curve closest to the $(0, 1)$ point**, and this method is usually called the **North-West corner or the Closest-to- $(0,1)$ criterion**.

Thus, individuals associated with a functional data $\chi \not\preceq \chi_{c_0}$ are classified as Affected (positive test), whereas individuals with $\chi \preceq \chi_{c_0}$ are classified as Non-Affected (negative test).

3. HOW DOES THE PROCEDURE WORK? A SIMULATION STUDY

3.1. Functional data samples

The sample curves χ_{ij} of the i -th group, $i = 1, 2$, are independent realizations of the continuous time process χ_i (functional biomarker defined in Affected ($i = 1$) and Non-Affected ($i = 2$) populations) defined as

$$\chi_i(t) = \mu_i(t) + \sum_{p=1}^{\infty} e^{-p/2} N_{pi} \psi_p(t); t \in T,$$

where $\mu_i(\cdot)$ are the mean functions, N_{pi} are iid random variables $N(0, \sigma_1)$, and $\psi_p(t) = 2^{1/2} \sin \{(p-1)\pi t\}$ ($p > 1$) and $\psi_1(t) \equiv 1$ are orthonormal basis functions. In addition, the functional data χ_{ij} are observed with error in a discrete set of points $t_1 < \dots < t_L$ of T , that is

$$X_{ijl} = \chi_{ij}(t_l) + \varepsilon_{ijl}, i = 1, 2, 1 \leq j \leq n_i, n = n_1 + n_2, 1 \leq l \leq L,$$

where, for each $i = 1, 2$, the errors ε_{ijl} are iid random variables $N(0, \sigma)$, with a standard deviation σ suitable to each situation.

Four families of mean functions μ_i and one type of error ε_{ijl} have been considered to set up different simulation scenarios (F1)-(F4). To implement the classification rule, it is needed to choose the representative curves M_A and M_{NA} among several centrality measures ($r.t$) and a semi-metric ($d.t$) (see Table 1).

Táboa 1: Types of representative curves $r.t$ and semi-metrics $d.t$ considered throughout of simulation study

Centrality Measures		Proximities between functional data	
$r.t$	Representative Curves	$d.t$	Semi-metric
1	(FMean) Mean curve	1	(L2) $L_2[T]$ metric
2	(depth.FM) Median with depth.FM	2	(Semi.Basis) semimetric.basis
3	(depth.mode) Median with depth.mode	3	(CCD) cosine correlation distance
5	(MMC) Minimum and maximum of data	4	(Semi.Deriv) semimetric.deriv
6	(depth.RP) Median with depth.RP	5	(Semi.PCA) semimetric.pca
7	(depth.RPD) Median with depth.RPD	6	(Semi.hshift) semimetric.hshift
8	(depth.FSD) Median with depth.FSD	7	(Semi.Fourier) semimetric.fourier
9	(depth.KFSD) Median with depth.KFSD	8	(Semi.mplsr) semimetric.mplsr

We consider a set of $R = 100$ replications, three sample sizes ($n_1 = n_2 = 20, 30, 50, 80$) and several values for the parameters σ and σ_1 in accordance to the simulation scenarios (F1)-(F4).

3.2. Results about discriminant power of test

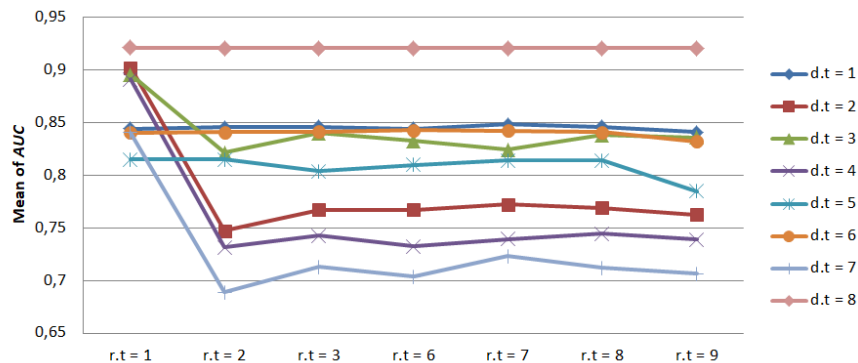


Figura 1: Means of area under ROC curve (AUC) for the several values of $r.t$ and $d.t$ when F1 scenario is considered. Sample size $n_1 = n_2 = 20$, $\sigma = 0,08$ and $\sigma_1 = 0,1$

3.3. Results about the predictive ability of test

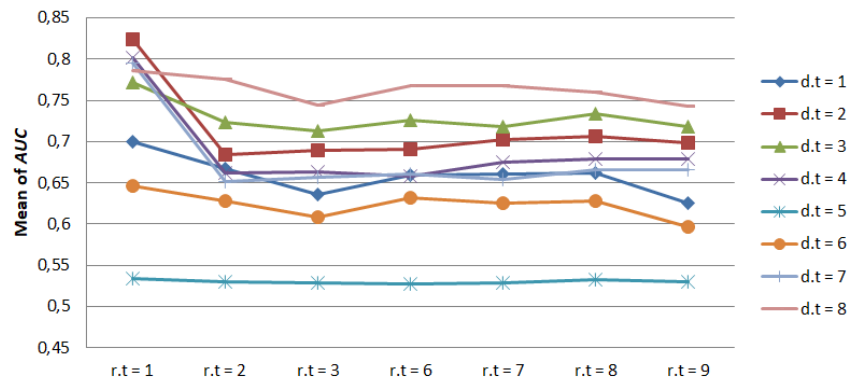


Figura 2: Means of area under ROC curve (AUC) for the several values of $r.t$ and $d.t$ when F2 scenario is considered. Sample size $n_1 = n_2 = 20$, $\sigma = 0,01$ and $\sigma_1 = 0,03$

Táboa 2: Means and standard deviations of the misclassification rates, obtained by classifying the 10 test samples, for scenarios F1 - F4

SCENARIO F1 (representative curve is: FMean r.t=1)				
	$n_1 = n_2 = 20$	$n_1 = n_2 = 30$	$n_1 = n_2 = 50$	$n_1 = n_2 = 80$
$\sigma = 0,04; \sigma_1 = 0,04$	$d.t = 8$ 2,75(4,227)	$d.t = 8$ 3,10(4,006)	$d.t = 8$ 2,30(3,512)	$d.t = 8$ 2,20(3,203)
$\sigma = 0,06; \sigma_1 = 0,06$	$d.t = 8$ 7,95(6,400)	$d.t = 8$ 7,20(6,085)	$d.t = 8$ 6,85(5,108)	$d.t = 8$ 6,60(5,811)
$\sigma = 0,08; \sigma_1 = 0,1$	$d.t = 8$ 16,30(7,234)	$d.t = 8$ 14,65(7,793)	$d.t = 8$ 14,75(7,013)	$d.t = 8$ 14,05(7,131)
SCENARIO F2 (representative curve is: FMean r.t=1)				
	$n_1 = n_2 = 20$	$n_1 = n_2 = 30$	$n_1 = n_2 = 50$	$n_1 = n_2 = 80$
$\sigma = 0,006; \sigma_1 = 0,008$	$d.t = 8$ 5,55(5,070)	$d.t = 8$ 5,40(5,350)	$d.t = 8$ 5,00(4,820)	$d.t = 8$ 5,15(5,100)
$\sigma = 0,01; \sigma_1 = 0,03$	$d.t = 8$ 19,5(10,237)	$d.t = 8$ 29,55(10,505)	$d.t = 8$ 26,95(9,531)	$d.t = 8$ 27,50(11,989)
$\sigma = 0,06; \sigma_1 = 0,06$	$d.t = 8$ 44,80(11,124)	$d.t = 8$ 45,36(21,323)	$d.t = 8$ 42,00(19,175)	$d.t = 8$ 35,65(18,878)
SCENARIO F3 (representative curve is: FMean r.t=1)				
	$n_1 = n_2 = 20$	$n_1 = n_2 = 30$	$n_1 = n_2 = 50$	$n_1 = n_2 = 80$
$\sigma = 0,5; \sigma_1 = 0,5$	$d.t = 7$ 1,95(3,616)	$d.t = 7$ 1,1(2,418)	$d.t = 1$ 0,75(2,057)	$d.t = 1$ 0,35(1,466)
$\sigma = 1,0; \sigma_1 = 1,0$	$d.t = 8$ 2,05(3,699)	$d.t = 8$ 1,75(3,436)	$d.t = 8$ 1,05(2,388)	$d.t = 8$ 0,65(1,966)
$\sigma = 1,5; \sigma_1 = 1,5$	$d.t = 8$ 4,75(4,994)	$d.t = 8$ 3,9(4,298)	$d.t = 8$ 3,55(3,846)	$d.t = 8$ 3,7(3,932)
SCENARIO F4 (representative curve is: FMean r.t=1)				
	$n_1 = n_2 = 20$	$n_1 = n_2 = 30$	$n_1 = n_2 = 50$	$n_1 = n_2 = 80$
$\sigma = 0,05; \sigma_1 = 0,05$	$d.t = 5$ 1,3(2,897)	$d.t = 5$ 1,4(3,104)	$d.t = 5$ 0,6(1,781)	$d.t = 5$ 0,3(1,193)
$\sigma = 0,075; \sigma_1 = 0,075$	$d.t = 8$ 1,9(3,161)	$d.t = 8$ 1,55(3,072)	$d.t = 8$ 1,4(3,022)	$d.t = 8$ 0,6(1,781)
$\sigma = 0,1; \sigma_1 = 0,1$	$d.t = 8$ 2,85(3,427)	$d.t = 8$ 2,3(3,289)	$d.t = 8$ 2,4(3,657)	$d.t = 8$ 1,95(3,170)

4. CONCLUSIONS

We have observed that functional diagnostic test proposed in this paper is a good option to discriminate two groups and classify new subjects when the biomarker is functional. We have considered four simulation scenarios whose mean functions present different degrees of separation and hence, different levels of difficulty to distinguish Affected subjects from Non-Affected ones. We have also considered diverse sample sizes and several levels of variability intra and inter curves, to cover a wider range of real situations, and we have evaluated multiple combinations for some parameters of the methodology.

REFERENCES

- Aneiros, G., Cao, R., Fraiman, R., Genest, C. and Vieu, P. (2019) Selected review on functional data analysis and related topics. *J. Multivariate Anal.*, 146, 1–6
- Cuevas, A. (2014) A partial overview of the theory of statistics with functional data. *J. Statist. Plann. Inference*, 147, 1–23
- Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics New York: Springer
- Horváth, L. and Kokoszka, P. (2012) *Inference for functional data with applications*. Springer
- Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F. and Alonzo, T.A. (2012) Extending induced ROC methodology to the functional context. *Biostatistics*, 13(4), 594–608
- Inácio, V., de Carvalho, M., Alonzo, T.A. and González-Manteiga, W. (2016) Functional Partial Area under the Curve Regression with an Application to Metabolic Syndrome Diagnosis. *Annals of Applied Statistics*, 10, 1472–1495

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

LA CALIDAD DE LA DEMOCRACIA: ¿ES EL GÉNERO DETERMINANTE PARA MEDIR LA CALIDAD DE LA DEMOCRACIA EN EL PERÚ?

González de la Rosa, L.A.¹, Sánchez-Martínez, L.J.²

¹ Departamento de Ciencia Política y de la Administración. Facultad de Ciencias Políticas y Sociología. Universidad Complutense de Madrid (UCM). Madrid.

² Departamento de Biodiversidad, Ecología y Evolución. Facultad de Biología. Universidad Complutense de Madrid (UCM). Madrid.

RESUMEN

La calidad de la democracia estudia la capacidad de un régimen político de maximizar su potencial. Desde la propuesta de Leonardo Morlino para medir la calidad de la democracia a partir de dimensiones referidas a los procedimientos, al contenido y al resultado, este trabajo realiza un análisis del nivel de igualdad para las mujeres y hombres en materia de educación, trabajo, salud y participación política desde el año 2007 al 2017 en el Perú. De este modo, se trata de ubicar en el análisis de la equidad de género. Para mostrar las diferencias en la distribución de logros entre mujeres y hombres en el Perú, es necesario ver la brecha de género en la dimensión propuesta por Morlino, ya que la equidad de género debería tener una presencia transversal en todas esas dimensiones. Los indicadores analizados en el presente estudio se dividen en 4 categorías de variables: Participación política, situación económica y laboral, salud y educación. La información acerca de las variables se ha extraído de la base de datos del XII Censo de Población, VII de Vivienda y III de Comunidades Indígenas del Perú. Mediante el análisis de componentes principales, el análisis de correlación y la elaboración de modelos mixtos de efectos aleatorios se ha estudiado primero la situación del 2017, y a partir de las diferencias encontradas, se ha estudiado si la variable sexo tiene una influencia significativa sobre el nivel de cada una de las variables analizadas, así como el impacto que ha tenido el transcurso del período de años 2007-2017 sobre las variables. Se ha comprobado que en la década estudiada, las mujeres en el Perú son las que encuentran más dificultades en todos los departamentos del país, pero también, que tanto hombres y mujeres están lejos de la situación ideal. Asimismo, se ha verificado que el departamento de residencia tiene impacto en los resultados tanto de hombres y de mujeres en todas las variables. Igualmente, son las mujeres de la sierra y de la selva las que presentan peores valores en casi todas las variables. Por ende, se ha demostrado que la desigualdad de género ha vulnerado la salud, el trabajo, la educación y la participación política de las mujeres en el Perú del 2007 al 2017 y que el esfuerzo del Estado peruano no ha sido suficiente para reducir las brechas entre mujeres y hombres. Por lo tanto, se concluye que la desigualdad de género es uno de los principales problemas de la democracia en el Perú, es decir, que el género es determinante para medir la calidad de la democracia en el Perú.

Palabras clave: Calidad de la democracia, enfoque de género, análisis multivariante, modelos mixtos.

1. INTRODUCCIÓN

¿Qué es la calidad de la democracia? ¿Cómo se mide la calidad de la democracia? ¿Existe un enfoque de género en la corriente principal de la Ciencia política que estudia la calidad de la democracia? ¿No es la igualdad un indicador principal de la medición de la calidad democrática? Si lo es, ¿en qué medida la igualdad de género es considerada un factor para evaluar la calidad de la democracia? ¿No es la mujer un agente de cambio vital en la sociedad para optimizar la calidad democrática de un país? ¿Cuán frágil es la democracia en el Perú? ¿Cuál es la situación de la democracia en el 2017? ¿Cómo ha evolucionado del 2007 al 2017? ¿Es el género determinante para medir la calidad de la democracia en el Perú? ¿Hasta qué punto las mujeres son ciudadanas que gozan de sus derechos plenamente en el Perú? ¿Son la educación, el trabajo, la salud y la participación política indicadores de igualdad? ¿Existen diferencias significativas entre hombres y mujeres en el Perú? Si las hay, ¿se distinguen por igual en la costa, sierra y selva?

En la actualidad, vivimos en una era democrática, donde la mayor parte de los países en América Latina, tienen procesos electorales, pluralismo político y participación política. Aunque seguimos teniendo gobiernos que abusan de sus poderes, desde la tercera ola democratizadora, en casi todos los Estados, las personas esperan que la democracia promueva la prosperidad económica y la igualdad social, viéndola

como el sistema que puede asegurar el bienestar de una nación. Sin embargo, a pesar de que la mayoría de los ciudadanos valoran el modelo democrático, no toman en consideración el papel que tienen en su funcionamiento más allá del voto. Lipset¹ explica que “los éxitos democráticos han reflejado las fuerzas variables de grupos políticos minoritarios y constelaciones afortunadas tanto o más que compromisos de los recientes gobernantes con el proceso democrático”. Por lo que es importante estudiar el éxito o fracaso de la democracia en la Ciencia Política, ya que depende tanto del ámbito político, como de los factores económicos y sociales que permiten que la calidad de la misma lleve a los países a ser evaluados. De esta manera, si bien se ha teorizado mucho sobre la democracia, el debate entre los alcances y los límites sobre la calidad de la democracia es aún pequeño. Dahl² ya distinguía la diferencia entre las democracias ideales y reales, pero difiere de país en país, tanto por su libertad e igualdad democrática como por su capacidad de toma y aplicación de decisiones.

Igualmente, del estudio de los diferentes enfoques de la calidad de la democracia que existen y que se complementan, se puede comprender que la calidad de la democracia es la capacidad de un sistema político de maximizar su potencial, y que no solo se inscribe dentro del ámbito político, sino que acapara los diversos espacios de una sociedad. Por lo tanto, la calidad de la democracia evalúa, como dice Lijphart³, el grado en el que un sistema cumple o no con la igualdad, la representación, la participación y la rendición de cuentas, o como dice Morlino⁵, la libertad y el control popular sobre las políticas públicas. La cuestión de la calidad de la democracia lleva a distinguir el rendimiento de los diversos regímenes políticos y verificar hasta qué punto los ciudadanos son respetados y considerados de la misma manera, tanto por las instituciones políticas como por la misma población, contribuyendo así con la legitimidad y la estabilidad política.

En ese sentido, es preciso estudiar la calidad de la democracia para las mujeres en América Latina porque se ven aún más debilitadas al ser la igualdad y la libertad prácticamente inexistentes para ellas, están subrepresentadas, se les excluye de los procesos de participación y de rendición de cuentas, y no se les toma en consideración para el control sobre las políticas públicas. Por consiguiente, analizar la calidad de democracia del Perú con un enfoque de género permite distinguir el rendimiento del régimen político peruano y verificar hasta qué punto las mujeres son respetadas como ciudadanas peruanas y consideradas de la misma manera, tanto por las instituciones políticas como por la misma población, contribuyendo así con la legitimidad y la estabilidad política del Perú.

Asimismo, resulta relevante abordar este complejo tema desde una perspectiva estadística que permita su operacionalización, ya que actualmente en la corriente principal de la Ciencia Política existe un déficit claro de estudios que empleen en sus análisis herramientas estadísticas, las cuales ofrecen una clara ventaja a la hora de evaluar y proporcionar resultados y conclusiones robustas que pueden tener impacto en el desarrollo de políticas futuras.

2. HIPÓTESIS Y OBJETIVOS

Los objetivos del presente estudio consisten en analizar la calidad de la democracia en el Perú para hombres y mujeres, y evaluar el nivel de implementación de la igualdad educativa, laboral, de salud y participación política, así como la capacidad de respuesta del Estado peruano a las necesidades de las mujeres en el período 2007-2017.

Nuestra hipótesis de partida es que existen diferencias significativas para la calidad de la democracia entre hombre y mujeres, más concretamente entre las tasas de las variables de trabajo, educación, salud y participación política. Además, entre los años del 2007 al 2017, la democracia en el Perú no ha logrado conseguir mejoras significativas. Más aún, se ha visto debilitada por el incumplimiento del Estado peruano de garantizar los derechos civiles y políticos, y los económicos, sociales y culturales de las mujeres

3. MATERIAL Y MÉTODOS

Los indicadores analizados en el presente estudio se dividen en 4 categorías de variables, tanto para hombres y mujeres, salvo en el caso de salud que solo es para mujeres debido a la naturaleza de las variables (p.e. Tasa de adolescentes embarazadas). Esta división responde a: la participación política, la situación económica y laboral, la salud y la educación. Las variables de educación son: la tasa de analfabetismo, la tasa neta de asistencia escolar, la tasa neta de matrícula escolar, la tasa de mujeres y hombres de 17 años de edad o más con al menos secundaria completa, y la tasa neta de matrícula a educación superior de mujeres y hombres de 17 a 24 años. Las variables de trabajo son: la tasa de egresados ocupados, la tasa de actividad económica, la tasa de informalidad, la tasa de trabajadores independientes, la tasa de personas sin ingresos propios y la relación empleo-población. Las variables de participación política son: la proporción de alcaldes, la proporción de congresistas y la proporción de regidores, y las variables de salud son: la tasa de mujeres y hombres con seguro de salud, la tasa de madres adolescentes o embarazadas por primera vez, la

tasa de atención prenatal por parte de profesionales de la salud, la tasa de partos atendidos en establecimientos de salud, la tasa de uso de métodos modernos de planificación familiar, la tasa de uso de métodos tradicionales de planificación familiar. Con estos indicadores se pretende abarcar tanto los niveles participación política y cultura política, como las libertades civiles y políticas, y los derechos socioeconómicos. Se han escogido las variables de educación, trabajo, salud y participación política porque son indicadores operacionalizados para poder medir la igualdad de acceso y ejercicio de los derechos y deberes expresados por la Constitución de la República del Perú, la equidad de género y el desarrollo humano. Asimismo, se han escogido por ser responsabilidad directa del Estado peruano. Del mismo modo, se emplea un método inductivo para tratar de ver en qué medida el dividir al Perú en las 3 regiones pueden demostrar las características de la población peruana, si hay una reproducción de la desigualdad en la misma escala para todas las mujeres y para comprender cuáles son los elementos a reivindicar para cuestionar y/o plantear si se puede hablar de una democracia de calidad.

La información acerca de las variables se ha extraído de la base de datos del Instituto Nacional de Estadística del Perú e Informática (INEI).

Dentro de las técnicas estadísticas empleadas durante la realización del estudio encontramos en primer lugar la aplicación de un análisis estadístico descriptivo (medidas de centralidad, mínimo, máximo, desviación estándar etc.). En un mismo sentido descriptivo se realizaron gráficos de estimación de densidad del núcleo (*Kernel density estimation plots*), así se pudo analizar el comportamiento de la función de densidad de los datos cuando se representaban una variable frente a otra.

Con el objetivo de analizar la situación actual de las mujeres en el Perú en el año 2017, se evaluó la interrelación de las variables analizadas realizándose un análisis de correlación entre cada una de ellas con el resto, para posteriormente presentar los resultados en forma de una matriz de correlaciones.

La correlación de Pearson (r) es un índice que mide el grado de covarianza entre distintas variables relacionadas linealmente, y sus valores fluctúan entre -1 y +1. De acuerdo con esto, la correlación entre dos variables es perfecta positiva (+1) cuando aumentan o disminuyen ambas variables de manera conjunta y a partir del cambio en una de ellas se puede predecir de forma exacta el cambio en la segunda, mientras que la relación es perfecta negativa (-1) cuando las variables se mueven en direcciones distintas, una variable disminuye y la otra aumenta. La ecuación matemática con la que se obtiene el coeficiente de correlación de Pearson es:

$$r_{XY} = \frac{\sum Z_X Z_Y}{N}$$

Por lo tanto, el coeficiente de correlación de Pearson (r_{XY}) se refiere a la media de los productos cruzados de los valores estandarizados de las variables X e Y ($Z_X Z_Y$). De esta manera, se ha escogido este método estadístico para comprobar la interdependencia entre las variables propuestas ya que se crean grupos de variables relacionadas entre sí, que pueden estar explicando fenómenos subyacentes que expliquen parte del comportamiento de los datos.

Con el objetivo de reducir la dimensionalidad de la serie de variables originales, las cuales entre sí presentan una correlación apreciable, se ha empleado la metodología de los análisis de componentes principales (ACP). En un primer nivel de análisis se manejaron los datos del año 2017 de mujeres por departamento, esto permitió analizar la situación actual de las mujeres discerniendo por su departamento de origen, y así evaluar que departamentos presentan una situación más óptima para las mujeres, tanto en educación, trabajo, participación política y salud. En una segunda fase del análisis se incorporaron los datos del año 2017 de hombres por departamento, para así poder analizar de forma comparativa la situación de ambos grupos, uno respecto al otro. Finalmente, partiendo de la situación anterior, se añadió un nuevo departamento ficticio en el cual los valores de sus variables eran aquellos que se establecieron como los ideales (p.e. 100% de Asistencia escolar, 0% de personas sin ingresos propios, 50% de proporción de alcaldes por sexos etc), con lo cual se pudo evaluar la situación real respecto a esta situación ideal, comparando a su vez los niveles establecidos de sexo y departamento.

Esta metodología es un tipo de análisis multivariante que crea nuevas variables no correladas entre sí (ortogonales), llamadas Componentes Principales (CPs), que se obtienen como resultado de crear combinaciones lineales ponderando las variables originales en ecuaciones del tipo:

$$CP_k = \lambda_{k1} X_1 + \lambda_{k2} X_2 + \dots + \lambda_{kp} X_p$$

Se trata de explicar la mayor proporción de la varianza total de las variables originales desde la menor cantidad posible de CPs, que representan de forma sintética y parsimoniosa la estructura y las interrelaciones existentes entre variables correlacionadas entre sí, las cuales expresan una misma dimensión del fenómeno medido⁵. Una vez se han extraído los CPs y se ha decidido el número de ellos que formarán parte del modelo, la matriz factorial muestra la relación entre las variables originales (filas) y los CPs (columnas) a través de los “*component loadings*” cuyos valores oscilan entre -1 y 1 y se interpretan como coeficientes de correlación, de acuerdo con la propiedad que cumplen los CPs de ser ortogonales y estar las variables estandarizadas. Los valores más altos (negativos o positivos) indican mayor relación entre la variable y el CP.

La modelización de cada una de las variables analizadas se realizó mediante un modelo lineal mixto de efectos aleatorios. Este procedimiento permitió, en primer lugar, valorar si el sexo se trata de un factor importante a la hora de determinar el valor de cada una de las variables por departamento, o si bien, ser hombre o mujer no resulta significativo a la hora de presentar valores más o menos altos en estas variables. En segundo lugar, se evaluó si el paso del tiempo podía influir sobre los valores de las variables, para ello se estableció el período de una década (datos del 2007 y datos del 2017) como período a incluir en el análisis. De esta manera, cada una de las variables se modelizó en base al sexo, al año y a la interacción entre ambos.

Estos modelos flexibilizan la regresión lineal ordinaria y permiten incorporar términos fijos y términos aleatorios. Muy a menudo, en las ciencias sociales se presentan variables respuesta que están relacionadas entre sí, ya sea por bloques, por medidas repetidas, o por otros factores que requieran de un modelo que introduzca efectos aleatorios⁶. Al igual que en los modelos GLM, la modelización se hace a través de una función de enlace que permite linealizar la variable respuesta y las covariables, mediante una transformación de la variable respuesta y obtener el predictor lineal condicionado⁷. El modelo empleado queda planteado como:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} \dots + \beta_n X_{nij} + \alpha_i + \varepsilon$$

El término aleatorio pertenece a los departamentos, los efectos fijos son: la variable de interés, que se modeliza en cada uno de los distintos modelos planteados, y las variables explicativas, que consisten en las variables de sexo y año.

Finalmente, la interpretación de los resultados está hecho con un enfoque de género porque el medir los niveles de hombres y mujeres en las variables propuestas necesita ser realizado a partir de las relaciones de género, ya que el género está vinculado a todos los aspectos de la vida política, social y económica, y representa una de las principales fuentes de desigualdad. Por lo que esta perspectiva permitirá examinar hasta qué punto hay un trato igualitario por parte del Estado peruano a sus ciudadanas y ciudadanos, y mostrará el carácter jerárquico entre géneros.

4. RESULTADOS

Dentro de los análisis estadísticos descriptivos realizados en el presente estudio encontramos los gráficos de densidad (*Kernel density estimation plots*) construidos entre las variables analizadas (datos del 2017). En la figura 1 se presenta el gráfico que relaciona las variables Sin ingresos y Analfabetismo. Los datos de mujeres aparecen representados en azul, mientras que los de los hombres en rojo. Como se aprecia, existe una clara diferenciación entre los datos de hombres y mujeres, los primeros se caracterizan por presentar una distribución más homogénea en cuanto a sus valores (nótese la curva de densidad), que se caracterizan por valores más bajos de Analfabetismo y Sin ingresos. Por otro lado, los datos de las mujeres presentan mucha más heterogeneidad en sus valores, caracterizándose por niveles más altos de Sin ingresos y también por una serie de departamentos que presentan elevadas tasas de Analfabetismo (superiores al 10%).

En la matriz de correlaciones se analiza la interrelación entre las variables analizadas de mujeres en el año 2017 (Figura 2). Se percibe la existencia de una alta correlación entre las variables de trabajadoras independientes ($r=0,87$) y de informalidad ($r=0,89$) con la de analfabetismo, es decir, en los departamentos con tasas más altas de analfabetismo hay más mujeres trabajando en el sector informal o siendo trabajadoras independientes. De la misma manera, la variable de relación empleo-población tiene una fuerte correlación con estas tres variables. ($r=0,78$; $r=0,79$; $r=0,70$). Lo que se infiere de estas correlaciones es que el comportamiento de estas variables está muy relacionado, por lo que en los departamentos donde estas variables de trabajo son más altas persisten la pobreza y la desigualdad por mantener altos niveles de precariedad, lo que mantiene a las mujeres en situación de marginación. De esta manera, las mujeres de

estos departamentos no pueden entrar al mercado laboral en igualdad de condiciones que las mujeres de los departamentos donde hay tasas más bajas porque carecen de la preparación necesaria. Además, la vulnerabilidad laboral a la que se enfrentan se comparte con el desconocimiento de sus derechos. Igualmente, según la matriz de correlaciones, las variables de trabajadoras independientes ($r=0,73$), informalidad ($r=0,69$) y analfabetismo ($r=0,71$) tienen una fuerte correlación con la tasa de mujeres sin ingresos propios. Esto significa que en los departamentos donde hay muchas mujeres que son trabajadoras independientes, que trabajan en la informalidad y son analfabetas también hay muchas mujeres sin ingresos propios. Por lo tanto, se deduce que las mujeres de departamentos con valores altos en estas variables tienen mucho más probabilidad de mantenerse en la pobreza al no tener autonomía económica.

En la figura 3 se presenta el análisis de componentes principales de los datos del año 2017 para mujeres y hombres por departamento, con respecto a la situación ideal. El CP1 explica el 44% de la variabilidad de los datos y el CP2 el 23%. En sentido positivo del CP1 se encuentran los hombres de la costa, sierra y selva, y las mujeres de la provincia de Lima, mientras que en sentido negativo se encuentran el resto de las mujeres de la costa, de la sierra y de la selva. En sentido positivo del CP2 se encuentran los hombres de la costa norte, de la selva y de la sierra, y en negativo los hombres de la costa sur y las mujeres de la costa. Las mujeres de la sierra y de la selva se encuentran dispersas en ambos sentidos del CP2. De este modo, son los hombres de la provincia de Lima, seguidos por los hombres de la costa sur y las mujeres de la provincia de Lima los que más cercanos están al punto ideal ya que poseen tasas más altas de egresados ocupados, de matrícula y asistencia escolar, de matriculación en educación superior, de finalización de educación secundaria y de logro académico. Sin embargo, lo que distingue a las mujeres de la provincia de Lima de los hombres de la sierra y de la selva son los altos porcentajes de congresistas, regidores y alcaldes de los segundos. Las más alejadas del punto ideal son las mujeres de la sierra y de la selva. A pesar de tener más acceso a seguro de salud que los hombres, predominan las altas tasas de informalidad, de relación empleo población, de analfabetismo, de mujeres sin ingresos propios, y las bajas tasas de las variables de participación política.

En la tabla 1 se presentan los resultados del modelo lineal mixto de efectos aleatorios que modeliza la tasa de Analfabetismo por departamento en función del sexo (hombre o mujer), del año (2007 o 2017) y de la interacción entre ambas variables.

Variables	Coefficiente	Error estándar	IC 95%	p-valor
Intercepto	4,800	0,996	(2,808; 6,792)	<0,000
Sexo (Mujer)	10,174	0,847	(8,480; 11,868)	<0,000
Año (2017)	-1,187	0,847	(-2,881; 0,507)	0,165
Sexo:Año	-3,159	1,198	(-5,555; -0,763)	0,010

Error absoluto medio: 2.168
R²: 0.756

Tabla 1: Modelización de la variable Analfabetismo en función del sexo, el año y la interacción entre ambos.

A la vista de los resultados del modelo mixto, la variable de sexo presenta un efecto significativo sobre nuestra variable respuesta, que se trata de la tasa de Analfabetismo (p-valor < 0,000), el valor del coeficiente estimado por el modelo (10,174) nos indica cuanto aumenta la tasa de Analfabetismo en los departamentos, cuando hablamos del grupo de mujeres frente al de hombres, es decir, en promedio la tasa de Analfabetismo de las mujeres de un departamento es 10,174 puntos más alta que la de los hombres de su mismo departamento, simplemente por el hecho de ser mujeres, con un IC 95% de: 8,480; 11,868 puntos.

En cuanto a la variable año, esta no presenta un efecto significativo sobre la variable respuesta (p-valor: 0,165), aunque el coeficiente estimado (-1,187) está indicando un descenso en la tasa de Analfabetismo desde los datos del año 2007 a los datos del año 2017. Este cambio no parece haber sido tan apreciable en la serie de datos, como se aprecia el intervalo de confianza incluye valores positivos (-2,881; 0,507).

Por otra parte, la interacción entre la variable sexo y año presenta unos valores significativos (p-valor: 0,010), lo cual nos está indicando que la variable respuesta de Analfabetismo no se ha visto afectada de la misma manera por la variable Año en el grupo de mujeres y en el grupo de hombres, ya que existe una interacción entre ambas. El coeficiente estimado por el modelo (-3,159) nos está indicando que la tasa de Analfabetismo en las mujeres tras el paso de los 10 años desciende en 3,159 puntos respecto al descenso que sufre la tasa de los hombres en ese mismo período de años, lo cual puede estar en relación con los

valores que presentaba la tasa de Analfabetismo para cada uno de los sexos en el 2007, 14,97% en mujeres y 4,78% en hombres, con lo cual es lógico que en las mujeres el descenso pueda ser mayor.

La tabla 2 presenta los resultados de la modelización de la variable Sin ingresos en función de las variables de sexo (hombre o mujer), año (2007 o 2017) y la interacción entre ambas variables.

Variables	Coefficiente	Error estándar	IC 95%	p- valor
Intercepto	12,629	0,857	(10,915; 14,343)	<0,000
Sexo (Mujer)	24,503	0,945	(22,613; 26,393)	<0,000
Año (2017)	-1,024	0,945	(-2,914; 0,866)	0,281
Sexo:Año	-5,334	1,336	(-8,006; -2,662)	<0,000

Error absoluto medio: 2.282

R²: 0.904

Tabla 2: Modelización de la variable Sin ingresos propios en función del sexo, el año y la interacción entre ambos.

Los resultados del modelo mixto, exponen un efecto significativo por parte de la variable sexo sobre la variable respuesta de Sin ingresos (p-valor < 0,000), el valor del coeficiente estimado por el modelo (24,503) nos está indicando cuanto aumenta la tasa de Sin ingresos en los departamentos, cuando nos referimos al grupo de mujeres frente al de hombres, es decir, en promedio las mujeres de un departamento presentan una tasa de Sin ingresos 24,503 puntos más alta que los hombres de su mismo departamento, con un IC 95% de: 22,613; 26,393 puntos.

En cuanto a la variable año, esta no presenta un efecto significativo sobre la variable respuesta de Sin ingresos (p-valor: 0,281), aunque el coeficiente estimado (-1,024) está indicando un pequeño descenso en la tasa de Sin ingresos desde los datos del año 2007 a los datos del año 2017, su intervalo de confianza incluye valores positivos (-2,914; 0,866).

En cuanto a la interacción entre las variables sexo y año, esta presenta un efecto significativo dentro del modelo (p-valor < 0,000), lo cual nos está indicando que la variable respuesta de Sin ingresos no se ha visto afectada de la misma manera por la variable Año en el grupo de mujeres y en el grupo de hombres, siendo, como está indicando el coeficiente estimado (-5,334) mayor el descenso en el grupo de las mujeres tras el paso de los 10 años respecto al descenso que sufre en los hombres en ese mismo periodo de años. Sin embargo, en el año 2007 la tasa media de Sin ingresos para las mujeres era del 37,13% mientras que para el grupo de hombres era mucho menor, del 12,62%, lo cual podría ser una razón para explicar porque entre las mujeres este descenso ha sido mucho más acusado.

5. CONCLUSIONES

Queda evidenciado una ausencia de enfoque de género en el análisis de la calidad de la democracia, ya que ningún índice o investigador se cuestiona la situación de las mujeres. Este trabajo ha demostrado que la desigualdad de género ha vulnerado la salud, el trabajo, la educación y la participación política de las mujeres en el Perú del 2007 al 2017. El esfuerzo del Estado peruano no ha sido suficiente para reducir las brechas entre los valores de los hombres y de las mujeres. Gracias al análisis de componentes principales, se ha verificado que las mujeres en el Perú encuentran más dificultades que los hombres en todos los departamentos del país, pero que tanto hombres y mujeres están lejos de la situación ideal.

Igualmente, la matriz de correlaciones ha permitido ver el comportamiento entre las variables seleccionadas, resaltando que la precariedad laboral y la falta de estudios van de la mano, así como un menor acceso a atención sanitaria.

De la misma manera, los resultados de los modelos mixtos de efectos aleatorios remarcan como la variable sexo tiene un peso significativo a la hora de explicar el comportamiento de las variables analizadas en los distintos departamentos, siendo las mujeres las que presentan una situación significativamente peor que los hombres. Se ha demostrado a su vez que los valores de las variables analizadas no han mejorado significativamente en el periodo 2007-2017.

Por consiguiente, se puede afirmar que la desigualdad de género es uno de los principales problemas de la democracia en el Perú, es decir, el género es determinante para medir la calidad de la democracia en el Perú. Lo expuesto en este trabajo pone en relieve como lo personal y lo político está conectado.

REFERENCIAS

1. Lipset, S. Reconsideración de los requisitos sociales de la democracia. *Este País*. 1995; 50: 2-18.
2. Dahl, R. *La Democracia: Una Guía para los ciudadanos*. Madrid: Taurus; 2004.
3. Lijphart, A. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. New Haven: Yale University Press; 2012.
4. Morlino, L. *La calidad de las democracias en América Latina: Informe para IDEA Internacional*. San José: Instituto Internacional para la Democracia y la Asistencia Electoral. 2014.
5. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. Boston: Pearson; 1989.
6. Bolker BM, Brooks ME, Connie JC, Geange SW, Poulsen JR, Stevens MHH et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*. 2008; 24(3): 127-35
7. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, linear, and mixed models*. New Jersey: Wiley; 2008.

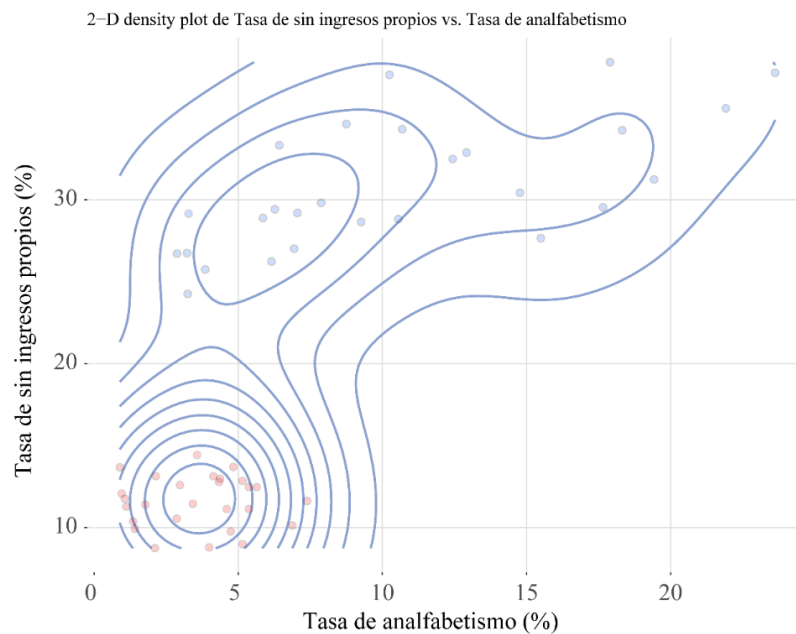


Figura 1: Gráfico que relaciona los valores de la variable Sin ingresos frente a la tasa de Analfabetismo en el año 2017. Los datos de mujeres se presentan en azul y los de los hombres en rojo, ambos para cada uno de los departamentos. Se presentan a su vez las curvas de densidad de los datos (*Kernel density estimation*).

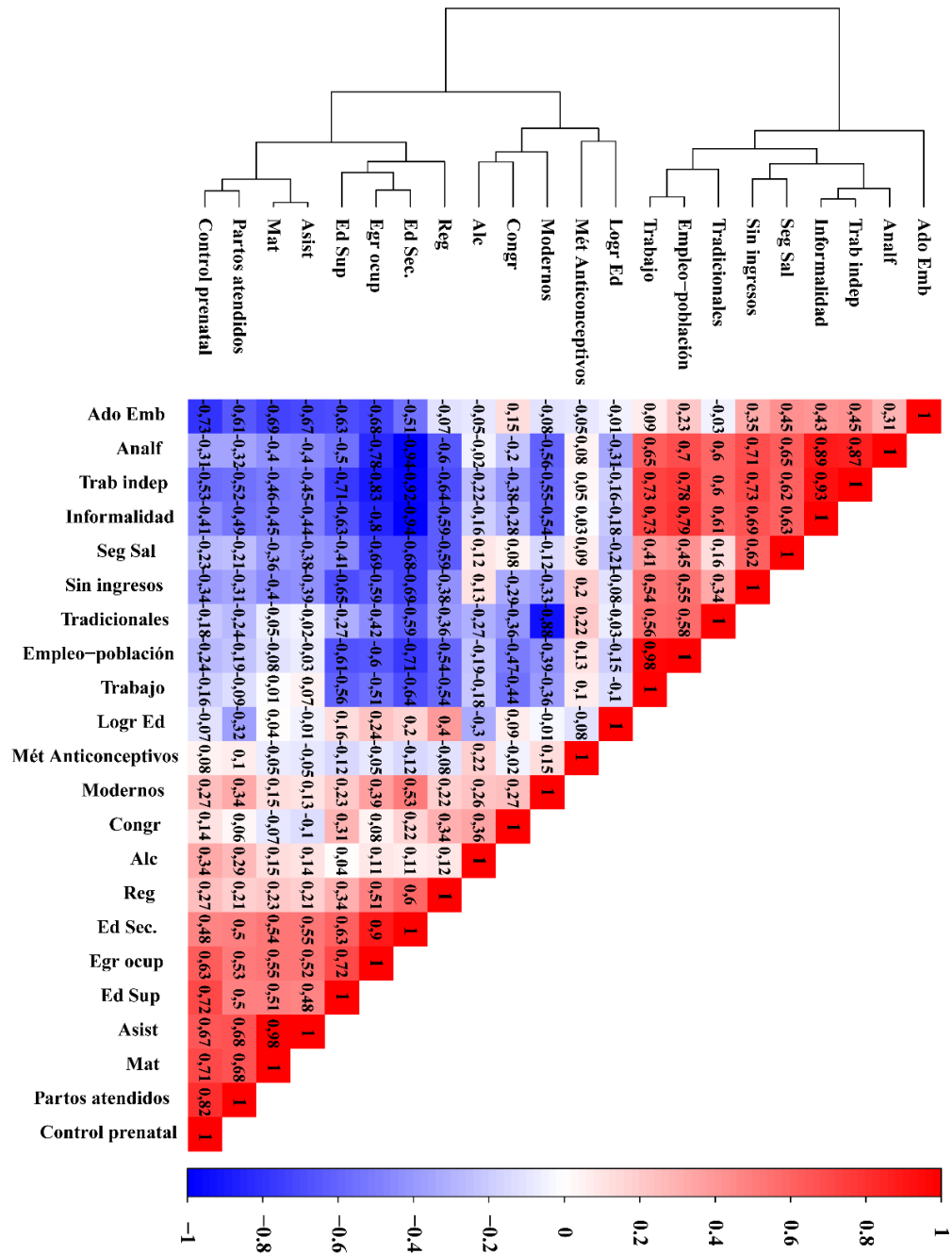


Figura 2: Matriz de correlaciones junto con la interrelación entre las variables analizadas de las mujeres en el año 2017.

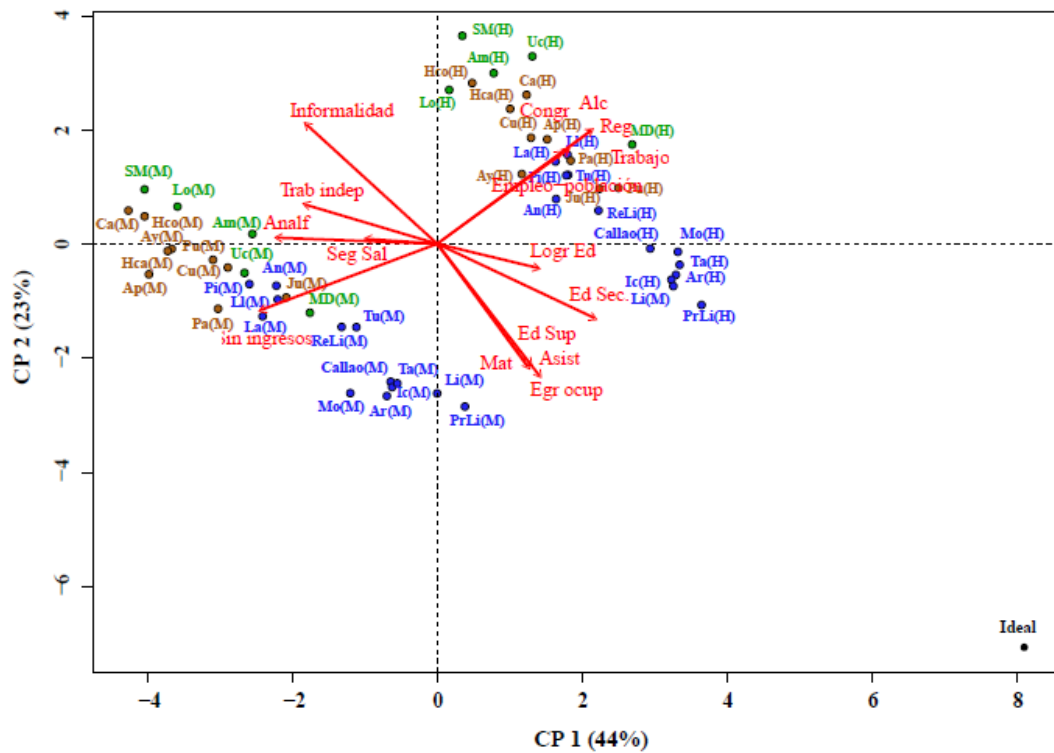


Figura 3: Análisis de componentes principales mostrando la situación de las mujeres (M) y de los hombres (H) en Perú, en base a las variables analizadas de educación, trabajo, participación política y salud en el año 2017, y frente a la situación ideal establecida. Los valores de los *component loadings* entre las variables y los CP se esquematizan con los vectores rojos.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

RAPOSa, UNA HERRAMIENTA GRATUITA PARA RESOLVER PROBLEMAS DE OPTIMIZACIÓN POLINÓMICA

Brais González Rodríguez¹, Julio González Díaz¹, Ángel Manuel González Rueda², Joaquín Ossorio Castillo³, Beatriz Pateiro López¹, Diego Rodríguez Martínez³, David Rodríguez Penas¹

¹Universidade de Santiago de Compostela

²Universidade da Coruña

³Instituto Tecnolóxico de Matemática Industrial

RESUMEN

En este artículo se presentará una nueva implementación de la *Reformulation Linearization Technique* (RLT) en el contexto de los problemas de optimización polinómica. El RLT es un algoritmo de ramificación y acotación basado en relajaciones lineales que asegura la convergencia a un óptimo global. Además de explicar la técnica RLT, se explicarán las diferentes mejoras que han sido añadidas a la implementación original, mejorando considerablemente el rendimiento del algoritmo. Finalmente, se mostrarán una serie de resultados comparando nuestra implementación con otras herramientas de referencia en el campo de la optimización.

Palabras y frases clave: optimización, polinómica, algoritmo, global.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

ESTIMATING THE RANDOM ARRIVAL RULE IN BANKRUPTCY: A CASE STUDY

Alejandro Saavedra-Nieves¹

¹SiDOR Research Group, Universidade de Vigo.

ABSTRACT

A bankruptcy problem (O'Neill, 1982) is an interactive situation in which several agents claim a portion of a good larger than the one available. These situations take the name from a very common problem in Economics as the bankruptcy of a company. In particular, the existence of several creditors that claim a fraction of a total estate is fundamental. Thomson (2015) provides a complete survey on this topic. A main goal in this class of problems consists of determining how we must divide the resource among all those agents who have claims on it. One of the most notable cases is the random arrival rule.

Although the random arrival rule has been analysed from a theoretical point of view in several papers (see Hwang (2015), for instance), its obtaining makes into a hard task when the number of involved agents enlarges. The main drawbacks concerning the random arrival rule for bankruptcy problems, similar to the ones described for obtaining the Shapley value of general TU-games, are computational. For large-scale problems, determining this rule substantially complicates since the complexity exponentially increases with the number of agents. In fact, Aziz (2013) provides an algorithm to determine it, but only for the special case of integer claims. Nevertheless, it does not solve the computational problem for the calculation in practice when considering large sets of real-valued claimants. Due to the wide applications of bankruptcy theory to real world, where exact solutions are often not possible in practice, sampling techniques (Cochran, 1977) become a useful alternative tool.

This work addresses a sampling methodology to estimate the random arrival rule in bankruptcy situations, based on simple random sampling with replacement. It is an adaptation of the procedure in Castro et al. (2009) for estimating the Shapley value of TU-games. It is especially useful when dealing with large-scale problems. This procedure is analysed by establishing the theoretical statistical properties and bounds for the incurred error. We evaluate this tool on examples for which this allocation rule can be exactly calculated. Finally, we apply this sampling method to provide a new system of quotas for the milk market in Galicia (Spain) after the suppression of the European regulation in 2015.

Key words and sentences: Bankruptcy, random arrival rule, sampling, game theory.

ACKNOWLEDGMENTS

Author acknowledges the financial support of *Ministerio de Economía y Competitividad* of the Spanish government under grant MTM2017-87197-C3-2-P and of Xunta de Galicia through the ERDF (*Grupos de Referencia Competitiva*) ED431C 2016-040.

REFERENCES

Aziz, H. (2013). Computation of the random arrival rule for bankruptcy problems. *Operations Research Letters*, 41(5), 499–502.

- Castro, J., Gómez, D., and Tejada, J (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5), 1726 - 1730.
- Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.
- Hwang, Y.-A. (2015). Two characterizations of the random arrival rule. *Economic Theory Bulletin*, 3(1), 43–52.
- O'Neill, B. (1982). A problem of rights arbitration from the Talmud. *Mathematical Social Sciences*, 2(4), 345–371.
- Thomson, W. (2015). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: an update. *Mathematical Social Sciences*, 74, 41–59.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

METODOLOGIAS BIOINFORMATICAS Y BIOESTADÍSTICAS EN EL ANÁLISIS DEL INBREEDING GENÓMICO

L. J. Sánchez-Martínez¹, C. L. Hernández¹, G. Pita², R. Calderón¹

¹Departamento de Biodiversidad, Ecología y Evolución. Facultad de Biología. Universidad Complutense de Madrid (UCM), Madrid

²Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid

RESUMEN

La ocurrencia de matrimonios consanguíneos ha sido secularmente una práctica habitual en muchas poblaciones humanas y las consecuencias biológicas en los descendientes de estas parejas emparentadas pueden ser importantes a nivel biomédico. La revolución genómica actual (genotipado masivo) ha supuesto un cambio de paradigma en el análisis del *inbreeding* en las poblaciones humanas actuales. En el presente estudio, se han caracterizado los patrones de *inbreeding* genómico de la metapoblación asentada en la región del Mediterráneo occidental. Se genotiparon 142 individuos autóctonos pertenecientes al sur de Iberia y Norte de África (Marruecos) mediante el chip Omni 2.5-8 (Illumina). Mediante los programas PLINK y R se estimaron los fragmentos *ROHs* y el coeficiente de *inbreeding* genómico (F_{ROH}). La información resultante fue combinada con otras poblaciones del Proyecto 1000 Genomas. Se aplicaron distintos análisis estadísticos de naturaleza multivariante, los cuales indicaron de una manera coincidente una más bien elevada homogeneidad en la estructura y niveles de *inbreeding* en las poblaciones ibéricas analizadas, aunque Andalucía occidental (Huelva) parece distinguirse moderadamente por sus tamaños totales y medios de fragmentos *ROHs* y coeficiente de *inbreeding* genómico, F_{ROH} . La población bereber de Marruecos refleja un acusado *inbreeding* reciente. Este escenario ha sido revelado por el alto peso de los fragmentos *ROHs* $> 5Mb$ dentro de su *pool* génico así como por un elevado F_{ROH} (0.0647). El modelo estadístico lineal mixto de efectos aleatorios construido indica que los fragmentos *ROHs* más grandes ($>5Mb$) son los que mejor explican el comportamiento del estadístico F_{ROH} a nivel del individuo.

Palabras clave: Consanguinidad, F_{ROH} , historia de la Población, Mediterráneo Occidental, Europa.

ABSTRACT

Consanguineous marriages have secularly been common in many worldwide human populations, and the biological consequences on the offspring of those biologically related couples can be relevant at a biomedical level. The current genomic revolution displayed through genome wide studies has led to a profound paradigm in the analysis of *inbreeding* in contemporaneous human populations. Here, we have characterized genomic *inbreeding* patterns and *inbreeding* levels of the metapopulation settled in the western edge of the Mediterranean region. One hundred and forty-two autochthonous individuals belonging to southern Iberia and North Africa (Morocco) were genotyped using the Omni 2.5-8 chip (Illumina). *ROHs* fragments and genomic *inbreeding* coefficient (F_{ROH}) were analyzed and estimated by using PLINK and R software. The obtained genomic information was combined with other populations drawn from 1000 Genomes Project. Different multivariate statistical analysis were applied, and interestingly Iberian populations showed a rather high homogeneity both in the structure and

inbreeding levels even though western Andalusia (Huelva) seems to be distinguished, in some way, by total and average sizes of *ROHs* fragments as well as by genomic inbreeding coefficient, F_{ROH} . The Berber population of Morocco presented signals of recent inbreeding, because of the high weight of large *ROHs* >5Mb within its gene pool and high F_{ROH} (0.0647). The mixed linear model of random effects built in the present work has unveiled that the largest *ROHs* fragments (>5Mb) represent the best-fit indicators for explaining the behavior of the F_{ROH} statistic at the individual level.

Keywords: Consanguinity, F_{ROH} , population History, Western Mediterranean, Europe.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

SMALL AREA ESTIMATION UNDER BIVARIATE NESTED ERROR REGRESSION MODELS

María Dolores Esteban¹, María José Lombardía², Esther López-Vizcaíno³, Domingo Morales¹,
Agustín Pérez¹.

¹Universidad Miguel Hernández de Elche

²Universidade da Coruña, CITIC

³Instituto Galego de Estatística

ABSTRACT

We introduce a bivariate nested error regression model and derive the empirical best linear unbiased predictors of means and a plug-in predictors of ratios. We approximate the mean squared error of the introduced predictors. We carry out simulation experiments to investigate the behavior of the residual maximum likelihood fitting algorithm, the predictors of domain means and ratios and the estimators of mean squared errors. Finally, we give an application to real data where the target is the small area estimation of averages and ratios of household annual expenditures in Spanish provinces.

Key words and sentences: Small area estimation, multivariate linear mixed models, best linear unbiased predictors, mean squared error estimation, household annual expenditures.

1. INTRODUCTION

The estimation of ratios in finite populations is usually done by estimating separately the numerator and the denominator with direct estimators. A direct estimator of the total or the average of a target variable in a domain uses only the data of that domain, it is basically unbiased with respect to the distribution of the sample design and its variance decreases when the sample size increases. The ratio estimators inherit part of these properties, so that both their biases and their variances also decrease as the sample size increases. However, domain sample sizes are typically small in the small area estimation setup. To avoid these problems, it is worthwhile to use area-level and unit-level multivariate statistical models because of the benefits of taking into account the correlation structure of target and auxiliary variables. See, for example, López-Vizcaíno et al. (2013, 2015) and Benavent et al. (2016).

We follow a multivariate approach for dealing with the small area estimation of expenditure means and ratios. More concretely, we propose fitting a bivariate nested error regression model to food and non-food household expenditures. This methodology improves the predictions of the domain expenditure parameters with respect to the separated modeling or the direct estimation approach. We develop predictors of domain means and ratios, give approximations to their mean squared errors (MSE), present some empirical investigations and illustrate the introduced methodology with an application to data from the Spanish Household Budget Survey (SHBS) of 2016.

2. THE BIVARIATE NESTED ERROR REGRESSION MODEL

Let U be a population of size N partitioned into D domains or areas $U_1, \dots, U_D$ of sizes $N_1, \dots, N_D$ respectively. Let $N = \sum_{j=1}^D N_d$ be the global population size. Let $y_{dj} = (y_{dj1}, y_{dj2})'$ be a vector of continuous variables measured on the sample unit j of domain d , $d = 1, \dots, D$, $j = 1, \dots, N_d$. For $k = 1, 2$, let $x_{dj k} = (x_{dj k1}, \dots, x_{dj k p_k})$ be a row vector containing p_k explanatory variables and let $X_{dj} = \text{diag}(x_{dj1}, x_{dj2})_{2 \times p}$ with $p = p_1 + p_2$. Let β_k be a column vector of size p_k

containing regression parameters and let $\beta = (\beta'_1, \beta'_2)'_{p \times 1}$. The population *bivariate nested error regression* (BNER) model is

$$y_{dj} = X_{dj}\beta + u_d + e_{dj}, \quad d = 1, \dots, D, j = 1, \dots, N_d, \quad (1)$$

where the vectors of random effects $u_d = (u_{d1}, u_{d2})'$ and the vectors of random errors $e_{dj} = (e_{dj1}, e_{dj2})'$ are all mutually independent with multivariate normal distributions

$$u_d \sim N_2(0, V_{ud}), \quad e_{dj} \sim N_2(0, V_{edj}), \quad d = 1, \dots, D, j = 1, \dots, N_d.$$

The 2×2 covariance matrices V_{ud} depend on 3 unknown parameters, $\theta_1 = \sigma_{u1}^2$, $\theta_2 = \sigma_{u2}^2$ and $\theta_3 = \rho_u$, and the 2×2 covariance matrices V_{edj} depend on 3 unknown parameters, $\theta_4 = \sigma_{e1}^2$, $\theta_5 = \sigma_{e2}^2$ and $\theta_6 = \rho_e$, i.e.

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & \rho_u \sigma_{u1} \sigma_{u2} \\ \rho_u \sigma_{u1} \sigma_{u2} & \sigma_{u2}^2 \end{pmatrix}, \quad V_{edj} = \begin{pmatrix} \sigma_{e1}^2 & \rho_e \sigma_{e1} \sigma_{e2} \\ \rho_e \sigma_{e1} \sigma_{e2} & \sigma_{e2}^2 \end{pmatrix}.$$

We implement a Fisher-scoring algorithm for estimating the residual maximum likelihood estimators of the model parameters.

Based on the BNER model (1), we derive the empirical best linear unbiased estimator, $\hat{\bar{Y}}_d^{eb}$, of $\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$ and the plug-in predictor $\hat{R}_d^{in}$ of the ratio $R_d = \bar{Y}_{d1}/(\bar{Y}_{d1} + \bar{Y}_{d2})$. We derive explicit-formula estimators of the corresponding MSEs and we carry out some simulations.

The new small area estimation methodology is applied to data from the Spanish household budget survey (SHBS) of 2016. The target is to estimate means of food and non-food household annual expenditures and ratios of household annual expenditures by Spanish provinces. For this sake, we fit a BNER model to the SHBS target data. The explanatory variables are

- Income. Total net annual household income (in euros).
- Family composition (FC). FC1: Single person or adult couple with at least one of the members being 65 years of age or older, FC2: Other compositions with a single person or a couple without children, FC3: Couple with children under 16 years old or adult with children less than or equal to 16 years old, FC4: Other households.
- Number of consumption units (multiplied by 10). $NCU = 10\{1 + 0.5(N_1 - 1) + 0.3N_2\}$, where N_1 is the number of people in the household aged 14 or older and N_2 is the number of people in the household under 14 years old.
- Rural. R1: Sparsely populated area, R0: Other areas.

A BNER model is selected and fitted to the data. After doing a model diagnostics, the predictors of the province target parameters are calculated and the corresponding MSEs are estimated. We show the benefits of using the introduced multivariate model-based methodology with respect to the univariate model-based or the direct approaches.

REFERENCES

- López-Vizcaíno, E., Lombardía, M.J., Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, 13, 2, 153-178.
- López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects.
- Benavent, R., Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372-390.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

UN NOVO CONTRASTE DE BONDADE DE AXUSTE PARA MODELOS DE REGRESIÓN CUANTIL CANDO A VARIABLE RESPOTA ESTÁ CENSURADA POLA DEREITA

Mercedes Conde Amboage¹, Ingrid Van Keilegom² e Wenceslao González Manteiga¹

¹Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela

² Research Centre for Operations Research and Business Statistics (ORSTAT). KU Leuven

RESUMO

Aínda que a regresión en media, axustada polo método de mínimos cadrados, acadou unha gran difusión no século XX, a regresión cuantil está recibindo unha atención crecente por parte dos investigadores en Estatística debido ás súas múltiples vantaxes. Destacamos que a regresión cuantil permite unha descrición máis detallada da distribución da variable resposta, non só na súa posición media ou central. Así mesmo, a regresión cuantil adáptase a situacións máis xerais que a regresión en media, sen necesidade de requirir normalidade da distribución do erro, goza de propiedades de robustez e permite abordar problemas de regresión con datos complexos (por exemplo, os datos censurados). Os modelos de regresión cuantil son moi adecuados para situacións clásicas de datos complexos, como a censura, o truncamento ou o sesgo de selección, porque evita os problemas derivados da falta de información, especialmente nos extremos das distribucións.

Dado que os modelos de regresión cuantil aplicados a datos censurados son moi empregados en diversos ámbitos como a Análise de Supervivencia ou a Biomedicina, xurde de forma natural a necesidade de verificar que dito modelo se axusta ben ao conxunto de datos co que se está a traballar. Aparece así a necesidade de presentar os contrastes de bondade de axuste no contexto da regresión cuantil que se abordarán ao longo deste traballo. Aínda que se poden atopar traballos sobre estimación de modelos de regresión cuantil baixo censura, os avances son moito menores no campo dos contrastes de bondade de axuste. Destaca a proposta de Wang (2008) que presenta un contraste dentro do contexto dos test baseados en métodos de suavizado estendendo as ideas de Zheng (1998).

Á vista do estado da arte, o obxectivo deste traballo será presentar un novo contraste de bondade de axuste que permita testear modelos paramétricos de regresión cuantil fronte a alternativas non paramétricas cando a variable resposta se encontra sometida a censura pola dereita. Dita nova proposta estará baseada na función de regresión integrada introducida por Stute (1997), de maneira que poidamos evitar o uso de parámetros de suavizado para definir o estatístico de contraste. É dicir, a idea sería estender o traballo de He e Zhu (2003) ao contexto dos datos censurados.

Analizarase o comportamento do novo contraste tanto baixo a hipótese nula como baixo a hipótese alternativa, concluíndo así a consistencia do mesmo. Ademais, unha vez que formulamos o estatístico de contraste, débese afrontar o problema de como levar a cabo o calibrado do test. Dada a complexidade de estimar as cantidades involucradas na representación asintótica do estatístico de contraste, propónse un procedemento bootstrap seguindo as ideas de Li e Datta (2001) para aproximar os valores críticos do novo contraste de bondade de axuste.

Por outra parte, preséntase un extenso estudo de simulación que mostra o bo comportamento da nova proposta. Concluíndo entón que o novo test mostra un bo axuste do nivel de significación independentemente do cuantil de interese, a porcentaxe de censura ou a distribución do erro considerados. Amais, é xeralmente máis potente que o

seu competidor natural. É importante destacar que a nova proposta tamén foi aplicada con éxito en contextos heterocedásticos.

Finalmente, o test proposto foi aplicado a un conxunto de datos reais mostrando deste xeito a súa utilidade de cara a validar modelos empregados no ámbito das Ciencias da Saúde.

Palabras e frases chave: regresión cuantil, datos censurados, contrastes de bondad de axuste, calibrado bootstrap.

REFERENCIAS

- He, X. e Zhu, L. X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, 98, 1013–1022.
- Li, G. e Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Annals of the Institute of Statistical Mathematics*, 53, 708–729.
- Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, 25, 613–641.
- Wang, L. (2008). Nonparametric test for checking lack of fit of the quantile regression model under random censoring. *Canadian Journal of Statistics*, 36, 321–336.
- Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, 14, 123–138.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

SPARSE SEMIPARAMETRIC REGRESSION FOR FUNCTIONAL DATA

Silvia Novo¹, Germán Aneiros^{1,2} and Philippe Vieu³

¹ MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña

² ITMATI

³ Institut de Mathématiques, Université Paul Sabatier

ABSTRACT

Dimensionality reduction in the regression context is addressed in the case when predictors are a mixture of functional variable and high-dimensional vector. A flexible model is proposed, which combines both sparse linear ideas together with semiparametrics. For this model, the asymptotic behaviour of the variable selection procedure which is proposed and the rates of convergence of the resultant estimators are provided. Some finite-sample simulated experiments and a real data application illustrate the usefulness of our methodology.

Keywords: Functional data analysis; Big data analysis; Variable selection; Sparse model; Dimension reduction; Semiparametrics.

1. INTRODUCTION

In many real problems, to predict the value of a random variable, observations of many other variables are available. In many fields of applications, data consisting of mixtures of functional and high-dimensional variables are common. In this situation, statistical methodologies have to cross both fields of Functional Data Analysis and Big Data. Accordingly, our purpose is to develop a regression model involving some scalar response, let's say Y , and predictors composed of some functional variable, $\mathcal{X}$, and some high-dimensional vector, $(X_1, \dots, X_{p_n})$. Since the parametric approach (see, for instance, Cardot and Sarda 2011) suffers from high lack of flexibility and the purely nonparametric methodology applied to functional data (see, for example, Ferraty and Vieu 2006) is difficult to interpret, semiparametrics could be a nice middle point between the two approaches.

Based on these considerations, the so-called semi-functional partial linear single-index model (SFPLSIM) is proposed. This model combines sparse linear ideas (see Aneiros et al. 2015 for details about the sparse semi-functional partial linear model), to control the high number of linear variables, p_n (which is allowed to go to infinity as n does), with functional semiparametric ideas for modelling the effect of the infinite-dimensional predictor, $\mathcal{X}$ (see, for instance, Ferraty et al. 2003, Ait-Saïdi et al. 2008 or Novo et al. 2019 for details about the single-index structure). Accordingly, a variable selection method and an estimate of the components of the model will be constructed, based on the penalized least-squares approach (PLS). Then, a wide set of asymptotic results will be derived. Finally, finite-sample behaviour of the method will be assessed through Monte Carlo experiments and an application to Tecator's data.

2. THE MODEL

The SFPLSIM is defined by the relationship

$$Y_i = X_{i1}\beta_{01} + \dots + X_{ip_n}\beta_{0p_n} + m(\langle \theta_0, \mathcal{X}_i \rangle) + \varepsilon_i, \quad (1)$$

$\forall i = 1, \dots, n$, where Y_i denotes a scalar response, $X_{i1}, \dots, X_{ip_n}$ are real random covariates and $\mathcal{X}_i$ is a functional random covariate valued in a separable Hilbert space $\mathcal{H}$ with inner product

$\langle \cdot, \cdot \rangle$. $\beta_0 = (\beta_{01}, \dots, \beta_{0p_n})^\top \in \mathbb{R}^{p_n}$ is a vector of unknown real parameters, $\theta_0 \in \mathcal{H}$ is an unknown functional direction and $m(\cdot)$ denotes an unknown smooth real-valued function. Finally, ε_i is the random error, which verifies $\mathbb{E}(\varepsilon_i | X_{i1}, \dots, X_{ip_n}, \mathcal{X}_i) = 0$.

3. THE VARIABLE SELECTION PROCEDURE

For the purpose of simultaneously estimating β -parameters and selecting relevant X -covariates in the SFPLSIM (1), we will follow the next steps:

First Step: We transform the SFPLSIM in a linear model by extracting from Y_i and X_{ij} ($j = 1, \dots, p_n$) the effect of the functional covariate $\mathcal{X}_i$ when is projected on the direction θ_0 . Specifically, denoting by $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip_n})^\top$, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, the fact that

$$Y_i - \mathbb{E}(Y_i | \langle \theta_0, \mathcal{X}_i \rangle) = (\mathbf{X}_i - \mathbb{E}(\mathbf{X}_i | \langle \theta_0, \mathcal{X}_i \rangle))^\top \beta_0 + \varepsilon_i, \quad \forall i = 1, \dots, n,$$

allows to consider the following approximate linear model

$$\tilde{\mathbf{Y}}_{\theta_0} \approx \tilde{\mathbf{X}}_{\theta_0} \beta_0 + \boldsymbol{\varepsilon}, \quad (2)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$. In expression (2), for any $(n \times q)$ -matrix $\mathbf{A}$ ($q \geq 1$), if $\mathbf{I}$ is the $(n \times n)$ -identity-matrix, we have denoted $\tilde{\mathbf{A}}_\theta = (\mathbf{I} - \mathbf{W}_{h,\theta}) \mathbf{A}$, where $\mathbf{W}_{h,\theta} = (w_{n,h,\theta}(\mathcal{X}_i, \mathcal{X}_j))_{i,j}$, with $w_{n,h,\theta}(\cdot, \cdot)$ being the weight function

$$w_{n,h,\theta}(\chi, \mathcal{X}_i) = \frac{K(d_\theta(\chi, \mathcal{X}_i)/h)}{\sum_{j=1}^n K(d_\theta(\chi, \mathcal{X}_j)/h)},$$

where $K : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a kernel function, $h > 0$ is a smoothing parameter and, for $\theta \in \mathcal{H}$, $d_\theta(\cdot, \cdot)$ is the semimetric defined as $d_\theta(\chi, \chi') = |\langle \theta, \chi - \chi' \rangle|$, $\forall \chi, \chi' \in \mathcal{H}$.

Second Step: The PLS approach is applied to model (2). Specifically, β_0 and θ_0 are estimated by considering a minimizer, $(\hat{\beta}_0, \hat{\theta}_0)$, of the penalized profile least-squares function

$$\mathcal{Q}(\beta, \theta) = \frac{1}{2} (\tilde{\mathbf{Y}}_\theta - \tilde{\mathbf{X}}_\theta \beta)^\top (\tilde{\mathbf{Y}}_\theta - \tilde{\mathbf{X}}_\theta \beta) + n \sum_{j=1}^{p_n} \mathcal{P}_{\lambda_{j_n}}(|\beta_j|),$$

where $\beta = (\beta_1, \dots, \beta_{p_n})^\top$, $\mathcal{P}_{\lambda_{j_n}}(\cdot)$ is a penalty function and $\lambda_{j_n} > 0$ is a tuning parameter.

Note that, simultaneously to the parameter estimation, the variable selection is carried out: if $\hat{\beta}_{0j}$ is a non-null component of $\hat{\beta}_0$, then X_j is selected as an influential variable.

4. CONCLUSIONS OF THEORY AND FINITE SAMPLE EXPERIMENTS

The existence of the penalized estimator is established as well as the corresponding rates of convergence. In particular, under some assumptions, we proved that there exists a local minimizer $(\hat{\beta}_0, \hat{\theta}_0)$ of $\mathcal{Q}(\beta, \theta)$ such that

$$\|\hat{\beta}_0 - \beta_0\| = O_p\left(\sqrt{s_n} \left(n^{-1/2} + \delta_n\right)\right) \text{ where } \delta_n = \max_{j \in S_n} \left\{ \left| \mathcal{P}'_{\lambda_{j_n}}(|\beta_{0j}|) \right| \right\}$$

where we have denoted the set of indices of the relevant variables and the cardinal of this set as $S_n = \{j = 1, \dots, p_n; \beta_{0j} \neq 0\}$ and $s_n = \#(S_n)$, respectively.

Rates of convergence for $\hat{\theta}_0$ and for the estimator of $m_{\theta_0} = m(\langle \theta_0, \mathcal{X}_i \rangle)$ are also derived (see Novo et al. “preprint”, for details). Furthermore, if we denote by $\hat{S}_n$ the estimation of S_n , that is $\hat{S}_n = \{j = 1, \dots, p_n; \hat{\beta}_{0j} \neq 0\}$, it is established that $\mathbb{P}(\hat{S}_n = S_n) \rightarrow 1$ as $n \rightarrow \infty$.

The simulation study and the application to real data will show the good behaviour of the presented method for finite sample size.

REFERENCES

- Ait-Saïdi, A., Ferraty, F., Kassa, R. and Vieu, P. (2008) Cross-Validated Estimations in the Single-Functional Index Model. *Statistics*, 42(6), 475–494.
- Aneiros, G., Ferraty, F. and Vieu, P. (2015) Variable selection in partial linear regression with functional covariate. *Statistics*, 49(6), 1322–1347.
- Cardot, H. and Sarda, P. (2011) Functional linear regression. *The Oxford handbook of functional data analysis*. Oxford University Press, 21–46.
- Ferraty, F., Peuch, A. and Vieu, P. (2003) “Modèle à Indice Fonctionnel Simple” (in French). *Comptes Rendus Mathématique de l’Académie des Sciences Paris*, 336(12), 1025–1028.
- Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Springer series in Statistics.
- Novo, S., Aneiros, G. and Vieu, P. (2019) Automatic and location-adaptive estimation in functional single-index regression. *Journal of Nonparametric Statistics*, 31(2), 364–392.
- Novo, S., Aneiros, G. and Vieu, P. Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. Preprint.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

BACKFITTING TESTS IN GENERALIZED STRUCTURED MODELS

Enno Mammen¹, Stefan Sperlich²

¹Institute for Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

²Geneva School of Economics and Management, University of Geneva, Bd du Pont d'Arve 40, 1211 Geneva, Suisse

ABSTRACT

We introduce bootstrap tests for semiparametric generalized structured models. They serve for testing different kinds of model specification like variable selection, separability, functional forms, the homogeneity of certain effects, etc., in a large class of semiparametric models. It is based on a comparison of semiparametric alternatives belonging to the same model class. Implementation and computational issues are discussed, and asymptotic theory is developed. Simulations and comparison studies show an excellent performance of the test procedures. Different applications illustrate usefulness and functioning of our method for the cases of generalised partial linear additive and varying coefficient models.

Key Words: Smooth Backfitting; Nonparametric Bootstrap Tests; Interactions; Separable Models.

1. INTRODUCTION

Generalised structured models (GSM) are an important tool in non- and semiparametric analysis. One can write this model class as

$$\Lambda(Y) = G\{Z, \beta, g(X)\} + S\{U, \gamma, s(T)\}\epsilon \quad (1)$$

with $E[\epsilon|Z, X, U, T] = 0$, Λ , G and S known (up to a finite dimensional parameter), β , γ unknown finite dimensional parameter, and $g(\cdot)$, $s(\cdot)$ nonparametric functions; Y is the dependent variable with observed covariates (Z, X, U, T) . GSMs allow for flexible modeling, they are easy to interpret, and there exist stable and reliable estimators. They avoid many of the problems we face in full dimensional nonparametric regression like the curse of dimensionality, difficulties of interpretation, and the lack of modeling possibilities. Yet, GSM do allow for the inclusion and modeling a large range of interactions, like synergy effects or the heterogeneity of effects. We concentrate on estimation methods that are numerically stable also for samples of small or moderate size, large dimensions or serious correlation among covariates.

Most common questions in empirical research are those about the separability of impacts and variable selection. Particularly interesting are then partial linear, additive and varying coefficient models, as well as combinations of them. This is provided by the so-called GSMs of Mammen and Nielsen (2003) which can be estimated by the smoothed backfitting (SB) method, first introduced by Mammen, Linton and Nielsen (1999). During the last 15 years various extensions have been proposed. A lot of emphasis has been put on the development of estimators for varying coefficient models, see Park, Mammen, Lee and Lee (2015) for a review.

So far, these methods are mainly used for some data explorative estimation; tools for further inference are still lacking. In contrast, various testing procedures were proposed for estimators based on marginal integration, see for example Yang, Park, Xue and Härdle (2006) for varying coefficient models, or Sperlich, Tjøstheim and Yang (2002) and Dette, von Lieres und Wilkau, and Sperlich (2005) for additive interaction models. The latter, however show that backfitting and marginal

integration are estimating essentially different things: while backfitting tries to optimally fit (or predict) the data, marginal integration tries to estimate only marginal impacts. Consequently, also the test procedures based on one or the other have quite different interpretations. Furthermore, for the existing (mostly marginal integration based) separability tests, they found poor power and serious problems regarding the error size when covariates are correlated. Among others, Nielsen and Sperlich (2005) found the same flaw for the classical backfitting procedure. The calibration problem of nonparametric tests where the null is semiparametric has been discussed in Sperlich (2014); but all other problems remain.

For all these reasons we propose procedures for specification testing in GSMs using SB. This completes an already rich literature on backfitting estimation for a large family of rather flexible models that we believe to be extremely useful in practice. For example, varying coefficients are a key instrument for studying the heterogeneity of causal effects (Sperlich and Theler, 2015). Our methods will be shown to work for testing significant interactions, variable selection, and constant coefficients (say, homogeneous effects); but we also discuss other structural hypotheses, including potential endogeneity. They are illustrated along quite different data examples. The first one was chosen for comparison reasons as these data have been studied intensively in papers on testing additivity: a nonparametric production model for predicting the output of farms in Wisconsin is tested for additivity. The second example studies in an macroeconomic growth model the homogeneity of effects. We can prove that the returns on human capital and investment on economic growth strongly vary over effect modifiers like size of middle class and inequality, a finding that facilitate explanations for some paradoxical empirical outcomes in the past. The third application tests for interaction of lymphocytes and plasma glucose concentration when predicting post-operational infection risk.

Technically spoken, our test procedure is based on the comparison of nonparametric estimators in L_2 -norm. A similar approach has been used in a series of testing problems where nonparametric estimators have been compared with parametric ones. A point that needs a lot of care in such tests are the bias terms of the nonparametric estimators as the test may else lose its omnibus character and one may arrive at an asymptotically linear tests that mainly checks for deviations from a parametric model in one direction, see e.g. Härdle and Mammen (1993). There exist different ways to avoid these bias effects. One way is to estimate the bias terms and to plug the estimated terms into the test statistic. Alternatively, one modifies the estimator in the null model such that it has asymptotically the same bias term as the nonparametric alternative. In the comparison of the two estimators the bias terms cancel out then. In this paper we follow the latter approach.

Another point that has been raised in the discussion of nonparametric L_2 -tests is the slow convergence of the test statistic to its normal limit. For this reason bootstrap procedures have been proposed that lead to more reliable distributional approximations and more accurate critical values. In the case of a parametric or a semiparametric hypothesis this can be easily implemented because one can resample from the null model. Note that our resampling method does not make use of any additional smoothing parameter. This is a quite nice feature compared to other related bootstrap procedures, cf. Sperlich (2014) for further discussion.

Besides the above mentioned reasons we use SB because it directly addresses the question if the model structure can approximately describe the data in the L_2 -optimal way. Furthermore, the assumptions for SB do not depend on the number of components but are comparable to lower-dimensional smoothing. As said above, in practice it turned out to perform excellent even in complex situations like strong correlation of covariates, sparse areas, high dimensions, etc., see Roca-Pardiñas and Sperlich (2010) and Nielsen and Sperlich (2005). Finally, the SB is well understood: basic asymptotic theory is given in Mammen, Linton and Nielsen (1999). Bandwidth choice and practical implementations are discussed in Mammen and Park (2005) and Nielsen and Sperlich (2005). SB methods for generalized additive models were introduced in Yu, Park and Mammen (2008). Linton and Mammen (2008) and Linton, Sperlich and Van Keilegom (2008) developed SB estimators for transformation models, i.e. in which a transformation lead to additive separability. Linton and Mammen (2002), Carrasco, Florens and Renault (2006), and Mammen and Yu (2009) outline that SB regression is in fact an example of a nonparametric model where the nonparametric function is given as a solution of an integral equation.

Related tests for variable selection are Härdle, Sperlich, and Spokoiny (2001) which use wavelets in additive models, or Sperlich, Tjøstheim, and Yang (2002), Yang, Sperlich, and Härdle (2003),

Härdle, Huet, Mammen, and Sperlich (2004) which propose bootstrap inference with marginal integration in generalized partial linear additive models. Gozalo and Linton (2001) consider the hypothesis of a generalized additive model and propose to use an L_2 -test statistic that compares a nonparametric additive estimator with a full dimensional kernel smoother.

REFERENCES

- Carrasco, M., J.P. Florens, and E. Renault, 2006, Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, Ed. J. Heckman und E. Leamer, Vol. 6. North Holland.
- Detle, H., C. von Lieres und Wilkau, and S. Sperlich, 2005, A comparison of different nonparametric methods for inference on additive models. *Journal of Nonparametric Statistics*, 17, 57-81.
- Gozalo, P. and O. Linton, 2001, A Nonparametric Test of Additivity in Generalized Nonparametric Regression with estimated parameters. *Journal of Econometrics*, 104, 1-48.
- Härdle, W., S. Huet, E. Mammen, and S. Sperlich, 2004, Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, 20, 265-300.
- Härdle, W. and E. Mammen, 1993, Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21, 1926-1947.
- Härdle, W., S. Sperlich, and V. Spokoiny, 2001, Structural tests in additive regression. *Journal of the American Statistical Association*, 96, 1333-1347.
- Linton, O., and E. Mammen, 2002, Nonparametric smoothing methods for a class of non-standard curve estimation problems. In: M.G. Akritas and D.N. Politis, (Eds.), *Recent Advances and Trends in Nonparametric Statistics*, Elsevier, Amsterdam.
- Linton, O., and E. Mammen, 2008, Nonparametric transformation to white noise. *Journal of Econometrics*, 142, 241-264.
- Linton, O., S. Sperlich, S., and I. Van Keilegom, 2008, Estimation of a Semiparametric Transformation Model. *Annals of Statistics*, 36, 686-718.
- Mammen, E., O.B. Linton, and J.P. Nielsen, 1999, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27, 1443-1490.
- Mammen, E., and J.O. Nielsen, 2003, Generalised structured models. *Biometrika*, 90, 551-566.
- Mammen, E., and B. Park, 2005, Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, 33, 1260-1294.
- Mammen, E., and K. Yu, 2009, Nonparametric estimation of noisy integral equations of the second kind. (with discussion) *J. Korean Statist. Soc.*, 38, 99-110.
- Nielsen, J.P., and S. Sperlich, 2005, Smooth backfitting in practice. *Journal of the Royal Statistical Society, B*, 67, 43-61.
- Park, B., E. Mammen, Y.K. Lee and E.R. Lee, 2015, Varying Coefficient Regression Models: A Review and New Developments. *International Statistical Review*, 83, 36-64.
- Roca-Pardiñas, J., and S. Sperlich, 2010, Feasible Estimation in Generalized Structured Models. *Statistics and Computing*, 20, 367-379.
- Sperlich, S. 2014, On the Choice of Regularization Parameters in Specification Testing: a critical discussion. *Empirical Economics*, 47, 427-450.
- Sperlich, S. and R. Theler, 2015, Modeling heterogeneity: a praise for varying-coefficient models in causal analysis. *Computational Statistics*, 30, 693-718.
- Sperlich, S., D. Tjøstheim, and L. Yang, 2002, Nonparametric Estimation and Testing of Interaction in Additive Models. *Econometric Theory*, 18, 197-251.
- Yang, L., B. Park, L. Xue and W. Härdle, 2006, Estimation and testing for varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association*, 101, 1212-1227.
- Yang, L., S. Sperlich and W. Härdle, 2003, Derivative Estimation and Testing in Generalized Additive Models. *Journal of Statistical Planning and Inference*, 115/2, 521-542.
- Yu, K., B. Park and E. Mammen, 2008, Smooth backfitting in generalized additive models. *Annals of Statistics*, 36, 228-260.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

IMPROVED Q -VALUES FOR DISCRETE AND HOMOGENEOUS TESTS

Marta Cousido-Rocha¹, Jacobo de Uña-Álvarez¹, Sebastian Döhler²

¹Department of Statistics and Operations Research, SiDOR Research Group, and Centro de Investigaciones Biomédicas (CINBIO), Spain

²CCSOR and Faculty of Mathematics and Science, University of Applied Sciences Darmstadt, Germany

ABSTRACT

Large scale discrete uniform homogeneous p -values arise in many applications, for example, in genome wide association studies. In this work several multiple testing procedures for such p -values are compared through simulations. Specifically, we consider the q -value approach based on five different estimators for the proportion of true null hypotheses π_0 . Two of them are classical estimators of π_0 for continuous p -values whereas the remaining three estimators are designed for discrete uniform homogeneous p -values. The q -value methods based on two of the latter estimators are novel in the multiple testing literature. The main simulated scenario is that of the two sample problem with low sample size and a large number of variables. Additional simulations based on the one sample problem scenario are also carried out. The main conclusion of this piece of work is that the discrete corrections may greatly improve the power of the q -values. An illustration with real data is included.

Keywords: Multiple testing procedures; Discrete p -values; High-dimensional data; Homogeneous p -values.

ACKNOWLEDGMENTS

This work has received financial support of the Call 2015 Grants for PhD contracts for training of doctors of MINECO, cofinanced by FEDER,UE (Ref. BES-2015-074958). The authors acknowledge support from MTM2014-55966-P (MINECO) and MTM2017-89422-P (MINECO/AEI/FEDER,UE) projects. We also acknowledge the financial support provided by “Xunta de Galicia” (ED431C 2016/040).

REFERENCES

- Chen, X., R. W. Doerge, and J. F. Heyse (2014). Methodology Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. arXiv:1410.4274v2 .
- Cousido-Rocha, M., J. de Uña-Álvarez, and S. Döhler (2019a). DiscreteQvalue: Improved q -values for discrete uniform and homogeneous tests. R package version 1.0.
- Cousido-Rocha, M., J. de Uña-Álvarez, and J. Hart (2019b). A two-sample test for the equality of univariate marginal distributions for high-dimensional data. Journal of Multivariate Analysis, <https://doi.org/10.1016/j.jmva.2019.104537>.
- Habiger, J. D. (2015). Multiple test functions and adjusted p -values for test statistics with discrete distributions. Journal of Statistical Planning and Inference 167, 1-13.
- Kulinskaya, E. and A. Lewin (2009). On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. Biometrika 96, 201-211.
- Liang, K. (2016). False discovery rate estimation for large scale homogeneous discrete p -values. Biometrics 72, 639-648.
- Storey, J. (2003). The positive false discovery rate: a bayesian interpretation and the q -value. The Annals of Statistics 31, 2013-2035.
- Storey, J. and R. Tibshirani (2003). Statistical significance for genomewide studies. Proceedings of National Academy of Science 100, 9440-9445.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

BANDWIDTH SELECTION FOR NONPARAMETRIC KERNEL PREDICTION

Inés Barbeito¹, Ricardo Cao² and Stefan Sperlich³

^{1,2}Group MODES, Department of Mathematics, CITIC, Universidade da Coruña

³Institute of Economics and Econometrics, University of Geneva

ABSTRACT

There exist many different methods to choose the bandwidth in kernel regression. If, however, the target is regression based prediction for samples or populations with potentially different distributions, then the existing methods can easily be suboptimal. This situation occurs for example in impact evaluation, data matching, or scenario simulations. We propose a bootstrap method to select a global bandwidth for nonparametric out-of-sample prediction. The asymptotic theory is developed, and simulation studies show the successful operation of our method. The method is used to predict nonparametrically the salary of Spanish women if they were paid along the same wage equation as men, given their own characteristics.

Keywords: Bandwidth Selection, Nonparametric prediction, Smooth Bootstrap.

1. INTRODUCTION

While there exist a considerable literature on bandwidth selection for kernel based nonparametric density and regression estimation, the problem of nonparametric prediction has largely been ignored. To our knowledge, such selection method does not exist albeit the relevance and frequency of such prediction problems in practise. They include for example any situation for which you want to predict counterfactuals like in impact evaluation (also known as treatment effect estimation). Other examples are statistical matching or data matching (see [1-3], and references therein), the imputation of missings (see e.g., [4-6], and references therein), or the simulation of scenarios. Note that we are not thinking of extrapolation far outside of the support of the observed covariates, a problem that would go beyond the here described ones, see [7]. We do not refer to bandwidth selection in stationary time series. In this context, various bandwidth and other model selection methods have been developed, see e.g. the review of Antoniadis, [8] or [9].

In all these situations have the following three features in common: you can think of a regression model with Y being the left-hand, and X the observed right-hand variables. You have one sample, denoted as 'source', in which both are given such that you can conduct a nonparametric regression. At the same time you have or simulate another sample or population, denoted as 'target', for which the same (as for 'source') potential response Y is not obtained. The basic assumption is that the dependence structure between, or in our case the conditional expectation of Y given X , $m(x) := E[Y|X = x]$ is the same in both populations. In data matching, and similarly when imputing missings, the Y were not sampled for the target sample; in scenarios the X of the target refer to an artificial, maybe future population, for which we just cannot observe any Y ; in counterfactual exercises you typically have Y observed for the target sample, but under a different situation, called 'treatment'. Then you use the source sample to impute the potential Y of the target group for the situation 'without treatment'. The difference between the observed Y (under treatment) and the imputed (without treatment) gives the so-called 'treatment effect for the treated'.

Our proposal relies on the so-called smooth bootstrap approach, see [10]. That is, you aim to draw bootstrap samples from a nonparametric pre-estimate of the joint distribution of (X, Y) . For the original source sample, and for each bootstrap sample you estimate $m(x)$. These allow us to

approximate the mean squared error of $\hat{m}(x)$ for any x inside the support of X . Finally you average these over the x_i observed in the target sample. We said 'you aim' because it can be shown that there exists a closed analytical form for the resulting MASE estimate. This simplifies the procedure drastically making it quite attractive in practise. One may argue that the exactness of this MASE approximation hinges a lot upon the pre-estimate. Yet, for finding the optimal bandwidth (or model) it suffices that our MASE approximations take their minimum at the same bandwidth as the true but unknown MASE. Our simulation studies show that this is actually the case.

2. THE BANDWIDTH SELECTION METHOD

Suppose we are provided with a complete sample $\{(x_i^0, y_i^0)\}_{i=1}^{n_0}$ from the source population with $X^0 \sim f^0$ and $m(x) := E_0[Y^0|X^0 = x]$. For the target population we only are provided with observations $\{x_i^1\}_{i=1}^{n_1}$ from density f^1 which is potentially different from f^0 . We are interested in predicting the expected $\{y_i^1\}_{i=1}^{n_1}$ assuming that $m(x) = E_1[Y^1|X^1 = x]$, or to estimate $\mathbb{E}[Y^1] = \mathbb{E}[m(X^1)]$. Moreover, if some 'outcomes' y_i are observed for the target population, their conditional expectation is supposed to differ from $m(\cdot)$; recall our example of outcome under treatment vs without, or see our application where $m(x)$ is the expected wage given x if you were a man.

For the prediction we have to estimate $m(\cdot)$ by a Nadaraya-Watson estimator $\hat{m}_h^{NW}$ with bandwidth h . Let us suppress for a moment the hyper-indices thinking for now always in the source sample with Y observed. The challenge is to find a bandwidth h which is MASE optimal for our predicting problem. The point-wise MSE, and afterwards the MASE are approximated by their bootstrap versions obtained as follows: Imagine $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ are bootstrap samples drawn from the kernel density $\hat{f}_g(x, y) = n^{-1} \sum_{i=1}^n K_g(x - X_i)K_g(y - Y_i)$ with bandwidth g . Then, for $\tilde{m}_h^{NW*}(x) f(x) = \hat{f}_h(x) \hat{m}_h^{NW}(x)$ we get

$$\tilde{m}_h^{NW*}(x) - \hat{m}_g^{NW}(x) = \frac{1}{n\hat{f}_g(x)} \sum_{i=1}^n K_h(x - X_i^*)(Y_i^* - \hat{m}_g^{NW}(x)), \quad (1)$$

where X^* has bootstrap marginal density $\hat{f}_g$, and $\mathbb{E}[Y^*|X^* = x] = \hat{m}_g^{NW}(x)$. Clearly, this is the bootstrap analogue to

$$\tilde{m}_h^{NW}(x) - m(x) = \frac{1}{nf(x)} \sum_{i=1}^n K_h(x - X_i)(Y_i - m(x)). \quad (2)$$

In order to compute the MASE we need to carefully distinguish between source and target sample, and have therefore to use the hyper-indices again. For finding a globally optimal bandwidth h we would like to minimise

$$MASE_{\tilde{m}_h^{NW}, X^1}(h) = \frac{1}{n_1} \sum_{j=1}^{n_1} \left[\mathbb{E}_0 \left[(\tilde{m}_h^{NW}(X_j^1) - m(X_j^1))^2 \right] \right], \quad (3)$$

where $\mathbb{E}_0$ refers to the expectation in the source population. We have in the bootstrap world

$$\begin{aligned} MASE_{\tilde{m}_h^{NW}, X^1}^*(h) &= \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{1}{\hat{f}_g^0(X_j^1)^2} \left[\left(1 - \frac{1}{n_0}\right) \cdot \left([K_h * \hat{q}_{X_j^1, g}^0](X_j^1)\right)^2 \right. \\ &\quad \left. + \frac{1}{n_0} \left[(K_h)^2 * \hat{p}_{X_j^1, g}^0\right](X_j^1) + \frac{g^2 d_K}{n_0^2} \sum_{i=1}^{n_0} [(K_h)^2 * K_g](X_j^1 - X_i^0) \right]. \end{aligned} \quad (4)$$

A bootstrap bandwidth selector for prediction is defined as

$$h_{BOOT}^{NW} = h_{MASE_{\tilde{m}_h^{NW}, X^1}^*} = \arg \min_{h>0} MASE_{\tilde{m}_h^{NW}, X^1}^*(h).$$

Note that the computation of h_{BOOT}^{NW} does not require the use of Monte Carlo approximation nor the nonparametric estimation of the density f^1 of the target population.

REFERENCES

- De Waal, T.; Pannekoek, J.; Scholtus, S. *Handbook of Statistical Data Editing and Imputation*; John Wiley: New York, NY, USA, 2011.
- Rässler, S. Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Aust. J. Stat.* **2004**, *33*, 153–171.
- Eurostat. *Statistical Matching: A Model Based Approach for Data Integration*; Methodologies and Working Papers; Eurostat: Luxembourg, 2013.
- Horton, N.J.; Lipsitz, S.R. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. *Am. Stat.* **2001**, *55*, 244–254.
- Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley: New York, NY, USA, 2004.
- Su, Y.-S.; Gelman, A.; Hill, J.; Yajima, M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.* **2010**, *45*, 1–31.
- Li, X.; Heckman, N.E. Local linear extrapolation. *J. Nonparametr. Stat.* **2003**, *15*, 565–578.
- Antoniadis, A.; Paparoditis, E.; Sapatinas, T. Bandwidth selection for functional time series prediction. *Stat. Probab. Lett.* **2009**, *79*, 733–740.
- Tschernig, R.; Yang, L. Nonparametric lag selection for time series. *J. Time Ser. Anal.* **2000**, *21*, 457–487.
- Cao-Abad, R.; González-Manteiga, W. Bootstrap methods in regression smoothing. *J. Nonparametr. Stat.* **1993**, *2*, 379–388.
- Barbeito, I.; Cao, R.; Sperlich, S. Bandwidth selection for nonparametric kernel prediction. Unpublished work, 2019.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

A TEST FOR COMPARING CONDITIONAL ROC CURVES WITH MULTIDIMENSIONAL COVARIATES

Arís Fanjul Hevia¹, Wenceslao González Manteiga², Juan Carlos Pardo Fernández³
and Ingrid Van Keilegom⁴

^{1,2}Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

³Departamento de Estatística e Investigación Operativa, Universidade de Vigo

⁴Research Centre for Operations Research and Business Statistics, KU Leuven

ABSTRACT

In any classification problem such as a diagnostic method – in which the aim is to discriminate the healthy from the diseased populations – one is concerned with minimizing the number of subjects that are misclassified. The Receiver Operating Characteristic (ROC) curves are commonly used in this context as a tool for studying the behaviour of the classification variables (see, for example, Pepe, 2003). They combine the notions of sensitivity (the ability of classifying a diseased patient as diseased) and specificity (the ability of classifying a healthy individual as healthy), two measurements that can be expressed in terms of the cumulative distribution functions of the diagnostic variables of the diseased and the healthy populations.

When there is more than one variable for diagnosing a certain disease one can compare their respective ROC curves in order to decide whether their discriminatory capability differs. There are several methodologies discussed in the literature for making that sort of comparisons (for a review of such methodologies, see Fanjul-Hevia and González-Manteiga, 2018), although most of them do not consider the possible effect that the presence of covariates can have in the performance of the test. It is important to take this information into account, because the diagnostic capability of a marker may change with the value of a covariate (Pardo-Fernández et al., 2014). One way of doing this is by using the *conditional ROC curve*.

Let Y^F and Y^G be the continuous diagnostic markers in the diseased and healthy population, respectively. Let $\mathbf{X} = (X_1, \dots, X_d)^t$ be a continuous d -dimensional covariate. $\mathbf{X}^F$ will represent the covariates in the diseased population, and $\mathbf{X}^G$ in the healthy population. Furthermore, let $F(y|\mathbf{x}) = P(Y^F \leq y | \mathbf{X}^F = \mathbf{x})$, and $G(y|\mathbf{x}) = P(Y^G \leq y | \mathbf{X}^G = \mathbf{x})$, where $\mathbf{x} = (x_1, \dots, x_d)^t \in R_{\mathbf{X}}$, with $R_{\mathbf{X}}$ being the intersection of the supports of $\mathbf{X}^F$ and $\mathbf{X}^G$. Then, given a fixed value $\mathbf{x} \in R_{\mathbf{X}}$, the conditional ROC curve is defined as

$$ROC^{\mathbf{x}}(p) = 1 - F(G^{-1}(1 - p|\mathbf{x})|\mathbf{x}), \quad p \in (0, 1).$$

By comparing the conditional ROC curves instead of the standard ROC curves one is able to incorporate the possible effect of the covariates in the analysis. A test for performing this comparison is proposed in Fanjul-Hevia et al. (2019) for the case of a continuous one-dimensional covariate. The objective here is to extend that methodology to the case in which we have a multidimensional covariate. Thus, the aim is to test, given a certain $\mathbf{x} \in R_{\mathbf{X}}$,

$$H_0 : ROC_1^{\mathbf{x}} = \dots = ROC_K^{\mathbf{x}}, \quad (1)$$

where K is the number of ROC curves that are being compared.

In order to do that, the problem is first transformed using projections in such a way that the multidimensional problem becomes a one-dimensional problem easier to

handle. This idea has been applied several times in the literature for reducing the dimension in goodness-of-fit problems (see, for example, Escanciano, 2006 or Cuesta-Albertos et al., 2019).

In this work, first it is shown how (1) can be transformed in a test with one-dimensional covariates using projections. Then, a methodology is proposed for testing that equivalent hypothesis. The results from a simulation study show the practical performance of the test in terms of level approximation and power. An application to real data is also presented to illustrate the procedure.

Keywords: bootstrap, covariates, hypothesis testing, projections, ROC curves.

REFERENCES

- Cuesta-Albertos J.A., García-Portugués E., Febrero-Bande M. and González-Manteiga W. (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, 47(1), 439–467.
- Escanciano J.C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6), 1030–1051
- Fanjul-Hevia A., González-Manteiga W. (2018). A comparative study of methods for testing the equality of two or more ROC curves. *Computational Statistics*, 33, 357–377.
- Fanjul-Hevia A., González-Manteiga W. and Pardo-Fernández J.C. (2019). A non parametric test for comparing conditional ROC curves. Manuscript submitted for publication.
- Pardo-Fernández J.C., Rodríguez-Álvarez M.X. and Van Keilegom I. (2014). A review on ROC curves in the presence of covariates. *REVSTAT- Statistical Journal*, 12, 21–41.
- Pepe M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DEFAULT IN CREDIT RISK

Rebeca Peláez Suárez¹, Ricardo Cao Abad², Juan Manuel Vilar Fernández²

¹Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain

²Research Group MODES, Department of Mathematics, CITIC, University of A Coruña and ITMATI, A Coruña, Spain

ABSTRACT

In this paper four nonparametric estimators of the probability of default in credit risk are proposed and compared. They are derived from estimators of the conditional survival function for censored data. Asymptotic expressions for the bias and the variance of these probability of default estimators are derived from similar properties for the conditional survival function estimators. A simulation study shows the performance of these four estimators. Finally, an empirical study based on modified real data illustrates their practical behaviour.

Keywords: risk analysis; probability of default; survival analysis; kernel method; local lineal fit.

1. INTRODUCTION

Once the credit scoring, assigned by a financial institution to a client who enjoys a personal credit, is known, it is interesting to find the probability that the borrower declares himself unable to face the debt contracted with the bank after some time (for example, one year) of its granting. The main aim of this work is to propose models to estimate this probability, known as the probability of default and denoted by PD .

For this purpose, the variable “time to default” will be studied, taking advantage of the information that the credit scoring provides about it. However, throughout the study of a set of credits, the default is not observed for all of them: some credits fall into default during the study, but some do not and others may be either cancelled or paid in advance. So, the variable “time to default” is censored.

On the other hand, the probability of default conditional to the credit scoring can be written as a transformation of the conditional survival function of the variable “time to default” by means of the following expression:

$$PD(t|x) = 1 - \frac{S(t+b|x)}{S(t|x)} \quad (1)$$

where $S(t|x)$ is the conditional survival function of time to default variable conditional to the credit scoring and b is the horizon of default.

As a consequence of both these features, censored data techniques and survival analysis can be used in this context and they provide estimators of the conditional survival function that derive in estimators of the probability of default through the previous transformation.

Three estimators for the conditional survival function have been first considered: Beran’s generalized product-limit estimator ([1]), weighted local linear (WLL) estimator ([2]) and Van Keilegom-Akritas estimator ([4]). They have been transformed to obtain the corresponding PD estimators.

2. NONPARAMETRIC PD ESTIMATORS

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) where X is the credit scoring, $Z = \min\{T, C\}$ is the observed maturity, T is the time to default, C is the time until the end of the study or the time until the anticipated cancellation of the credit and $\delta = I_{\{T \leq C\}}$ is the uncensoring indicator.

The Beran estimator's expression, which is the generalization of the Kaplan-Meier estimator to the case in which a covariate is involved, is the following:

$$\hat{S}_h^B(t|x) = \prod_{i=1}^n \left(1 - \frac{I_{\{Z_i \leq t, \delta_i=1\}} w_{i,n}(x)}{1 - \sum_{j=1}^n I_{\{Z_j < Z_i\}} w_{n,j}(x)} \right), \quad (2)$$

where the weights are

$$w_{i,n}(x) = \frac{K((x - X_i)/h)}{\sum_{j=1}^n K((x - X_j)/h)}, \quad i = 1, \dots, n,$$

where K is a kernel function (typically a density function to be picked up by the user) and $h > 0$ is a smoothing parameter.

The weighted local linear estimator is based on the following nonparametric regression model

$$V_t = \Psi_t(T) = m(X) + \varepsilon, \quad (3)$$

where $V_t = \Psi_t(T) = I_{\{T > t\}}$ to get the regression function estimator to be an estimator of the conditional survival function, $\hat{m}_t(x) = \hat{S}(t|X = x)$, for a fixed value of t .

Let $\{(X_{[i]}, Z_{(i)}, \delta_{[i]})\}_{i=1}^n$ be a simple random sample which is sorted according to the values $\{Z_i\}_{i=1}^n$ of the population (X, Z, δ) where $X_{[i]}$, $\delta_{[i]}$ are the concomitants of $\{Z_i\}_{i=1}^n$ and consider

$$\begin{aligned} S_{n,l}(x) &= \sum_{i=1}^n (X_{[i]} - x)^l w_{[i],h} W_{[i],n}, \\ T_{n,t}(x) &= \sum_{i=1}^n \Psi(Z_{(i)}) (X_{[i]} - x)^l w_{[i],h} W_{[i],n}, \end{aligned}$$

for $l = 0, 1, 2$, where $w_{[i],h} = K_h(X_{[i]} - x)$ with K a kernel function, $K_h(u) = K(u/h)/h$ and $h = h_n$ a smoothing parameter, are the covariate weights and

$$W_{[i],n} = \frac{\delta_{[i]}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{[j]}},$$

are the Kaplan-Meier censoring weights. The weighted local linear regression estimator (WLL) proposed in [2] is given by

$$\hat{S}_h^{WLL}(t|x) = \hat{m}_h^{WLL}(x) = \frac{S_{n,2}(x)T_{n,0}(x) - S_{n,1}(x)T_{n,1}(x)}{S_{n,2}(x)S_{n,0}(x) - S_{n,1}^2(x)}. \quad (4)$$

A problem in the WLL estimator was detected. It is not designed to estimate a survival function that has to be non-increasing and whose range lies in the interval $[0, 1]$, so, using it in its original version, PD estimations that take values greater than one or negative are obtained. As an alternative to this estimator, a local constant estimator with Nadaraya-Watson weights (WNW) is proposed. This estimator respects the isotony and the belonging to the interval $[0, 1]$ of the conditional survival function, returning correct estimations of the PD . The expression of this survival estimator is given by

$$\hat{S}_h^{WNW}(t|x) = \hat{m}_h^{WNW}(x) = \frac{\sum_{i=1}^n \Psi(Z_{(i)}) w_{[i],h} W_{[i],n}}{\sum_{i=1}^n w_{[i],h} W_{[i],n}}, \quad (5)$$

In [4] and [5] a nonparametric estimator of the conditional survival function is proposed. In order to define this estimator, the following nonparametric regression model is assumed:

$$T = m(X) + \sigma(X)\varepsilon,$$

where $m(x) = E(T|X = x)$ is the unknown regression curve; $\sigma(x)$ is the conditional standard deviation, enabling a possible heterocedastic model, and ε is the error variable. Taking into account that

$$F(t|x) = F_\varepsilon\left(\frac{t - m(x)}{\sigma(x)}\right),$$

obtaining consistent estimator of $\sigma(x)$ and $m(x)$ and estimating F_ε with the Kaplan-Meier estimator based on the regression residuals, an estimator of the conditional survival function, $\hat{S}_h^{VKA}(t|x)$, is available.

Now, replacing $S(t|x)$ with the conditional survival estimators $\hat{S}_h^B(t|x)$, $\hat{S}_h^{WLL}(t|x)$, $\hat{S}_h^{WNW}(t|x)$ and $\hat{S}_h^{VKA}(t|x)$ in (1), the four nonparametric estimators of $PD(t|x)$ are obtained. They are denoted by $\widehat{PD}_h^B(t|x)$, $\widehat{PD}_h^{WLL}(t|x)$, $\widehat{PD}_h^{WNW}(t|x)$ and $\widehat{PD}_h^{VKA}(t|x)$.

3. ASYMPTOTIC RESULTS

Under certain conditions and assuming that expressions for bias and variance of conditional survival function estimator are available, a theoretical result has been proven. It allows to obtain asymptotic properties of the bias and the variance of the corresponding PD estimator. This is the case of the WLL and Van Keilegom-Akritas survival estimators for which asymptotic expressions of bias and variance are known ([2], [4]). Thus it was possible to get results for the WLL and Van Keilegom-Akritas estimators of the probability of default. The asymptotic expressions of bias and variance for the Beran estimator of the PD are known and were proven along with the asymptotic normality of the estimator in [3]. Nevertheless, these are complex expressions that make it difficult to obtain an approximation of the MSE or MISE since they depend on several unknown parameters.

4. SIMULATION STUDY

In order to analyse the behaviour of the four PD estimators by simulation, several models have been considered. In the first studies carried out, Beran estimator provides the best results because it gives the best approximations (smaller mean squared error) and less computation time. The WNW estimator reduces the mean squared error committed by WLL. The WLL and Van Keilegom-Akritas estimators for the distribution function turn out to be better than Beran's with heavy censoring. A simulation study was carried out to check whether this property is inherited by the corresponding PD estimators. The results show that the Van Keilegom-Akritas estimator produces a smaller mean squared error in the right tail of the "time to default" distribution when the censoring is heavy. These four estimators involve a smoothing parameter in the covariate (credit scoring); the choice of this bandwidth parameter was examined in the simulation study.

Another fact observed in the simulation is that the PD estimators conditional to the credit scoring present excessive variability even using a large bandwidth to smooth the covariate. Therefore, a modification of the proposed estimators based on smoothing in the time variable has been considered. The first analyses seem to indicate that these estimators have a better behaviour.

References

- [1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report*, University of California.
- [2] Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. *Recent Advances and Trends in Nonparametric Statistics*, Michael G. Akritas and Dimitris N. Politis, pp. 217–231.
- [3] Cao, R., Vilar, J.M. and Devia, A. (2009). Modelling consumer credit risk via survival analysis (with discussion) *Statistics and Operations Research Transactions*, **33**, pp. 3–30.

- [4] Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics*, **27**:5, pp. 1745–1784.
- [5] Van Keilegom, I., Akritas, M.G. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics and Data Analysis*, pp. 487–500.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

ESTIMACIÓN NO PARAMÉTRICA DE MÁXIMA VEROSIMILITUD DE LA SUPERVIVENCIA LIBRE DE ENFERMEDAD CON DATOS DE CORTE TRANSVERSAL

Beatriz Piñeiro Lamas¹ y Jacobo de Uña Álvarez^{1,2}

¹Grupo de investigación SiDOR & CINBIO, Universidade de Vigo

²Departamento de Estadística e Investigación Operativa, Universidade de Vigo

RESUMEN

La supervivencia libre de enfermedad es una medida de pronóstico adecuada cuando el paciente puede sufrir una recaída o una complicación a lo largo de su evolución clínica. La estimación de la supervivencia libre de enfermedad a partir de datos de corte transversal no es trivial, ya que incluye problemas de censura y truncamiento sobre distintas variables aleatorias. En este trabajo se propone un estimador de la supervivencia libre de enfermedad basado en la máxima verosimilitud no paramétrica, y se introduce un algoritmo iterativo para su cómputo. Las propiedades del estimador se estudian mediante simulaciones. La principal conclusión de nuestra investigación es que el estimador propuesto tiene un buen comportamiento, siendo superior a otros estimadores existentes. Como ilustración, aplicamos el estimador de máxima verosimilitud al estudio de la evolución de pacientes en Unidades de Cuidados Intensivos.

Palabras y frases clave: Análisis de supervivencia, Censura, Kaplan-Meier, Modelos multi-estado, Supervivencia bivalente, Truncamiento.

1. INTRODUCCIÓN

En el Análisis de Supervivencia el interés se centra típicamente en el tiempo que transcurre hasta un evento final, que suele ser la muerte. En ocasiones, sin embargo, la variable de interés es el tiempo libre de enfermedad o evento, que indica el tiempo que transcurre hasta un evento intermedio (como puede ser una complicación post-operatoria o la recaída de una enfermedad, si ocurre) o hasta la muerte (si el evento intermedio no ocurre).

En los últimos años se ha incrementado el uso de medidas alternativas a la supervivencia total para el estudio y comparación de nuevos tratamientos o terapias. Una de esas alternativas es el tiempo libre de progresión (*progression-free survival*). La progresión de la enfermedad suele provocar síntomas que afectan a la calidad de vida de los pacientes, por lo que retrasarla es muy relevante. Recolectar datos de supervivencia total requiere en general un tiempo de seguimiento largo. Además, es común que un paciente reciba distintos tratamientos a lo largo de su vida. Si el interés se centra el comportamiento de un determinado tratamiento y consideramos el tiempo de supervivencia total, puede que en dicho tiempo influyan los tratamientos que el paciente haya recibido con posterioridad al tratamiento de interés. Por esta razón el tiempo libre de progresión, que es un tiempo libre de evento en el cual el evento intermedio es la progresión de la enfermedad, se está usando cada vez más en ensayos clínicos de cáncer (Buyse et al., 2007; Korn y Crowley, 2013).

En este trabajo, a modo de ilustración, consideramos los datos de un estudio llevado a cabo en Unidades de Cuidados Intensivos (UCIs) en Israel entre 1999 y 2000. Para más detalles sobre estos datos, consultar Vakulenko-Lagun y Mandel (2016), Vakulenko-Lagun et al. (2017) o de Uña-Álvarez y Mandel (2018). En este caso el evento intermedio de interés es BSI (*bloodstream infection*), una infección severa que es una de las principales causas de muerte en los hospitales.

Como evento final se considera la salida de la UCI, ya sea debida a la muerte del paciente o a su alta hospitalaria. En la Figura 1 se muestra el diagrama multi-estado correspondiente, que consta de tres estados y tres posibles transiciones. El estado 1 representa la ausencia de BSI, y desde él se puede transitar al estado 2 (presencia de BSI) o al 3 (alta o muerte). También es posible transitar desde el estado 2 al estado 3. Una dificultad que presenta el análisis de los datos de las UCIs es que fueron obtenidos mediante corte transversal, incluyéndose en la muestra únicamente los pacientes que estaban en la UCI en una fecha determinada, realizándose posteriormente un seguimiento de cada caso hasta completar 30 días desde el ingreso. El muestreo por corte transversal provoca un truncamiento en los datos que, en ausencia de adecuadas correcciones, puede llevar a estimaciones sesgadas de la supervivencia.

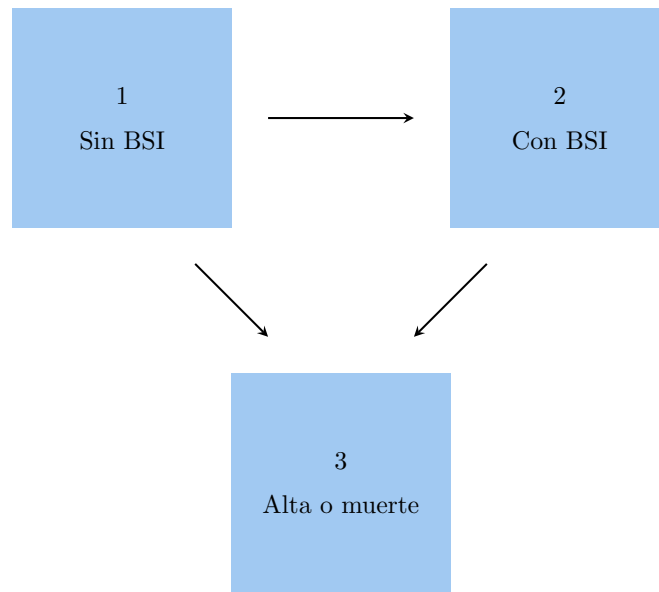


Figura 1: Gráfico del modelo enfermedad-muerte progresivo (modelo multi-estado) para los datos de las UCIs.

Tales muestreos de corte transversal, en los que únicamente se observan los individuos que se encuentran en progreso (esto es, vivos) en la fecha de corte, se realizan con frecuencia en el contexto del Análisis de Supervivencia debido a su simplicidad (Wang 1991; Fluss et al. 2013). En muestreo de corte transversal, denotamos la información muestral por $(L_i, Z_i, \delta_i, T_i, \Delta_i)$, $1 \leq i \leq n$. La variable L denota el tiempo desde el inicio del proceso (ingreso hospitalario, por ejemplo) hasta la fecha de corte transversal. Sea Z^0 el tiempo hasta el primero entre los eventos intermedio y final, y sea T^0 el tiempo hasta el evento final. Durante el seguimiento de los individuos puede haber censura por la derecha debido a distintas causas (pérdidas durante el seguimiento o fin de estudio, por ejemplo); sea C el tiempo de censura potencial. Consideramos, sin pérdida de generalidad, que $P(L \leq C) = 1$. En lugar de observar (L, Z^0, T^0) se observa $(L, Z, \delta, T, \Delta)$ con $L \leq T$, donde $Z = \min(Z^0, C)$, $\delta = I(Z^0 \leq C)$, $T = \min(T^0, C)$ y $\Delta = I(T^0 \leq C)$. Asumimos que $(L_i, Z_i, \delta_i, T_i, \Delta_i)$, $1 \leq i \leq n$, son copias iid de la variable $(L, Z, \delta, T, \Delta)$ condicionada por $L \leq T$ (equivalentemente, por $L \leq T^0$). La variable L juega entonces el papel de variable de truncamiento por la izquierda sobre la variable T^0 . La Figura 2 ilustra esta situación de muestreo por corte transversal; en esta Figura se puede ver que es posible tener información de individuos con Z^0 menor que el tiempo de truncamiento L , siempre que éste sea menor que T^0 .

La estimación no paramétrica de la función de supervivencia de T^0 bajo truncamiento por la izquierda y censura por la derecha ha sido ampliamente investigada en la literatura estadística (Tsai et al., 1987; Gijbels y Wang, 1993; Zhou y Yip, 1999). Más reciente es, sin embargo, la investigación sobre la estimación de la supervivencia libre de enfermedad con datos de corte trans-

versal (de Uña-Álvarez, 2017). La principal dificultad en este contexto es que se quiere reconstruir la distribución de una variable sobre la cual el truncamiento actúa de una manera indirecta. Se requiere por tanto el diseño de nuevos estimadores.

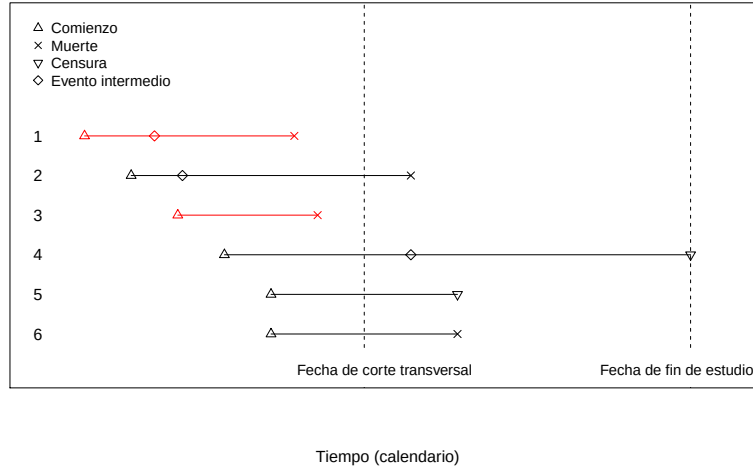


Figura 2: Tiempos de supervivencia con truncamiento por la izquierda y censura por la derecha de un modelo de corte transversal con un evento intermedio. Los segmentos rojos no se observan.

En la Sección 2 se introduce un nuevo estimador no paramétrico de la supervivencia libre de enfermedad para datos de corte transversal mediante un criterio de máxima verosimilitud, y se propone un algoritmo iterativo para su cómputo. El estimador introducido maximiza la verosimilitud condicional de los tiempos de supervivencia libres de enfermedad observados (quizá censurados) (Z_i, δ_i) , dados los tiempos de truncamiento L_i , siendo una generalización natural del estimador Kaplan-Meier a nuestro contexto. En la Sección 3 se estudian las propiedades del estimador de máxima verosimilitud (EMV) mediante simulaciones. El estudio de simulación sirve además para comparar el EMV con otros dos posibles estimadores de la supervivencia libre de enfermedad: el estimador límite-producto integral (ELPI) y el estimador basado en el factor de reducción (EFR) (de Uña Álvarez, 2017). En la Sección 4 se ilustra el EMV mediante su aplicación a los datos de las UCIs. En la Sección 5 damos las conclusiones principales de este trabajo.

2. ESTIMADOR DE MÁXIMA VEROSIMILITUD

Consideremos el modelo de truncamiento-censura basado en las hipótesis siguientes:

- (A1) $P(L \leq C) = 1$
- (A2) (L, C) independiente de (Z^0, T^0)

La hipótesis de independencia (A2) indica que los procesos de truncamiento y censura no están relacionados con el proceso de interés, lo cual es realista en la práctica. Bajo (A2), la condición (A1) no es restrictiva ya que, de no cumplirse, el par (L, C) puede sustituirse por otro par (L^*, C^*) que tenga la distribución de (L, C) condicionada por $L \leq C$.

Proposición 1. Bajo el modelo (A1)–(A2) la verosimilitud de los (Z_i, δ_i) condicionada a los tiempos de truncamiento L_i es proporcional a

$$\mathcal{L}_c \propto \prod_{i=1}^n \frac{P(Z^0 = Z_i, L_i \leq T^0)^{\delta_i} P(Z^0 > Z_i, L_i \leq T^0)^{1-\delta_i}}{P(T^0 \geq L_i)}. \quad (1)$$

Demostración. Para demostrar el resultado calculamos bajo (A1)–(A2) la aportación de cada dato (Z_i, δ_i) a la función de verosimilitud condicionada, dependiendo del valor de su indicador de censura:

Caso 1. Si $\delta_i = 1$:

$$P(Z = Z_i, \delta = 1 | L = L_i, L \leq T) = \frac{P(L = L_i, C \geq Z_i)P(Z^0 = Z_i, T^0 \geq L_i)}{P(L_i \leq T^0)P(L = L_i)}.$$

Caso 2. Si $\delta_i = 0$:

$$P(Z = Z_i, \delta = 0 | L = L_i, L \leq T) = \frac{P(L = L_i, C = Z_i)P(Z^0 > Z_i, T^0 \geq L_i)}{P(L_i \leq T^0)P(L = L_i)}.$$

Suponiendo que la distribución conjunta de (L, C) no está relacionada con la de (Z^0, T^0) , llegamos al resultado anunciado. $\square$

Nos centraremos en la clase de distribuciones discretas bivariantes con puntos masa (Z_i, T_i) , $1 \leq i \leq n$, siendo $w_i = P(Z^0 = Z_i, T^0 = T_i)$. De esta forma, suponiendo que no hay empates entre los Z_i , la verosimilitud condicional se puede reescribir como

$$\mathcal{L}_c \propto \prod_{i=1}^n \frac{w_i^{\delta_i} [\sum_{j=1}^n w_j I(Z_j > Z_i)]^{1-\delta_i}}{\sum_{j=1}^n w_j I(T_j \geq L_i)}, \quad (2)$$

donde para el segundo factor en el numerador hemos utilizado que $\delta_i = 0$ implica $T_i = Z_i$.

Definimos el EMV de la función de distribución de Z^0 , $F_{Z^0}(z) = P(Z^0 \leq z)$, como $\hat{F}_{Z^0}(z) = \sum_{i=1}^n \hat{w}_i I(Z_i \leq z)$, donde $\hat{w} = (\hat{w}_1, \dots, \hat{w}_n)$ es el maximizador de (2) sujeto a $w_i \geq 0$, $1 \leq i \leq n$, y $\sum_{i=1}^n w_i = 1$. La función de supervivencia libre de enfermedad $S_{Z^0}(z) = P(Z^0 > z)$ se estima entonces por $\hat{S}_{Z^0}(z) = 1 - \sum_{i=1}^n \hat{w}_i I(Z_i \leq z)$. Nótese que $\hat{F}_{Z^0}(z)$ es la primera marginal de la conjunta $\hat{F}_{Z^0 T^0}(z, t) = \sum_{i=1}^n \hat{w}_i I(Z_i \leq z, T_i \leq t)$.

El estimador $\hat{S}_{Z^0}(z)$ puede verse como una generalización del estimador de Kaplan-Meier a nuestro contexto de muestreo por corte transversal. De hecho, cuando no hay truncamiento ($L_i = 0$, $1 \leq i \leq n$), el denominador en la ecuación (2) es igual a uno y se obtiene la verosimilitud clásica para datos censurados por la derecha. Por otra parte, cuando no hay censura ($\delta_i = \Delta_i = 1$, $1 \leq i \leq n$), $\hat{S}_{Z^0}(z)$ pondera Z_i con el salto del estimador límite-producto de $F_{T^0}(t) = P(T^0 \leq t)$ (Tsai et al., 1987) en el i -ésimo concomitante T_i . En estos dos casos particulares, por tanto, el EMV tiene fórmula explícita. Para su cómputo en la situación más general de censura y truncamiento propondremos a continuación un algoritmo iterativo. Para $k \in \{1, \dots, n\}$ se tiene que

$$\frac{\partial \log(\mathcal{L}_c)}{\partial w_k} = 0 \iff w_k = \frac{\delta_k}{\sum_{i=1}^n \frac{I(L_i \leq T_k)}{\sum_{j=1}^n w_j I(L_i \leq T_j)} - \sum_{i=1}^n (1 - \delta_i) \frac{I(Z_k > Z_i)}{\sum_{j=1}^n w_j I(Z_j > Z_i)}}; \quad (3)$$

por tanto,

(a) si $\delta_k = 0$ entonces $w_k = 0$; esto es, los individuos censurados reciben un peso nulo;

(b) si $\delta_k = 1$ entonces $w_k = \left[\sum_{i=1}^n \frac{I(L_i \leq T_k)}{\sum_{j=1}^n w_j I(L_i \leq T_j)} - \sum_{i=1}^n (1 - \delta_i) \frac{I(Z_k > Z_i)}{\sum_{j=1}^n w_j I(Z_j > Z_i)} \right]^{-1}$.

Para hallar $(w_1, \dots, w_n)$ se propone entonces el siguiente algoritmo, siendo n_u el número de datos no censurados e $iter$ el número máximo de iteraciones:

Paso 0 (solución inicial):

$$w_k^{(0,*)} = \frac{1}{n_u} \text{ para los individuos tales que } \delta_k = 1;$$

$$w_k^{(0,*)} = 0 \text{ para los individuos tales que } \delta_k = 0.$$

Paso $l \in \{1, \dots, iter - 1\}$:

$$w_k^{(l)} = \left[\sum_{i=1}^n \frac{I(L_i \leq T_k)}{\sum_{j=1}^n w_j^{(l-1,*)} I(L_i \leq T_j)} - \sum_{i=1}^n (1 - \delta_i) \frac{I(Z_k > Z_i)}{\sum_{j=1}^n w_j^{(l-1,*)} I(Z_j > Z_i)} \right]^{-1} \text{ si } \delta_k = 1;$$

$$w_k^{(l)} = 0 \text{ si } \delta_k = 0;$$

$$\text{Normalización: } w_k^{(l,*)} = \frac{w_k^{(l)}}{\sum_{j=1}^n w_j^{(l)}}.$$

El algoritmo se detiene cuando la diferencia máxima entre la curva de supervivencia estimada en la iteración $l - 1$ y curva de supervivencia estimada en la iteración l sea menor que cierto umbral de error, en nuestro caso fijado en 10^{-5} . Además se permite un número máximo de $iter = 150$ iteraciones.

La restricción $\sum_{i=1}^n w_i = 1$ en el proceso de maximización de (2), que motiva el paso de normalización del algoritmo, implica de alguna manera la estimación de una función de distribución de probabilidad condicionada. Actualmente, si el mayor de los Z_i es censurado, el estimador Kaplan-Meier ordinario es una distribución impropia, y lo mismo cabría esperar para el EMV con datos de corte transversal. Para permitir la posibilidad $\sum_{i=1}^n w_i < 1$, el paso de normalización se obvia en el cómputo de la solución final.

Una propiedad interesante de la verosimilitud condicional (1) es que puede justificarse también bajo el modelo de truncamiento-censura alternativo

$$(B1) \ P(L \leq C) = 1$$

$$(B2) \ L \text{ independiente de } (Z^0, T^0)$$

$$(B3) \ C - L \text{ independiente de } (L, Z^0, T^0) \text{ dado } L \leq T^0$$

El siguiente resultado establece formalmente tal propiedad. Este modelo (B1)–(B3) ha sido ampliamente utilizado en la literatura sobre estimación con datos de corte transversal (de Uña-Álvarez, 2017). Sin embargo, la condición (B3) no es válida para los datos de las UCIs (al tener C un punto masa en tiempo 30 días), que sí se adaptan no obstante al modelo (A1)–(A2) discutido arriba.

Proposición 2. Bajo el modelo (B1)–(B3) la verosimilitud de los (Z_i, δ_i) condicionada a los tiempos de truncamiento L_i es proporcional a (1).

Demostración. Análogamente a la demostración de la Proposición 1, calcularemos la contribución a la verosimilitud condicional de cada dato (Z_i, δ_i) bajo (B1)–(B3) distinguiendo los casos $\delta_i = 1$ y $\delta_i = 0$. En lo sucesivo R^* denota la función de supervivencia de $C - L$ condicionada a $L \leq T$.

Caso 1. Si $\delta_i = 1$:

$$\begin{aligned}
 P(Z = Z_i, \delta = 1 | L = L_i, L \leq T) &= \frac{P(Z = Z_i, \delta = 1, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= \frac{P(Z^0 = Z_i, Z_i - L_i \leq C - L, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= \frac{P(Z_i - L_i \leq C - L | L \leq T) P(Z^0 = Z_i, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= R^*(Z_i - L_i) \frac{P(Z^0 = Z_i, L = L_i, L_i \leq T^0)}{P(L = L_i, L_i \leq T^0)} \\
 &= R^*(Z_i - L_i) \frac{P(Z^0 = Z_i, L_i \leq T^0)}{P(L_i \leq T^0)}.
 \end{aligned}$$

Caso 2. Si $\delta_i = 0$:

$$\begin{aligned}
 P(Z = Z_i, \delta = 0 | L = L_i, L \leq T) &= \frac{P(Z = Z_i, \delta = 0, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= \frac{P(C - L = Z_i - L_i, Z_i < Z^0, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= dR^*(Z_i - L_i) \frac{P(Z_i < Z^0, L = L_i | L \leq T)}{P(L = L_i | L \leq T)} \\
 &= dR^*(Z_i - L_i) \frac{P(Z_i < Z^0, L = L_i, L_i \leq T^0)}{P(L = L_i, L_i \leq T^0)} \\
 &= dR^*(Z_i - L_i) \frac{P(Z_i < Z^0, L_i \leq T^0)}{P(L_i \leq T^0)}.
 \end{aligned}$$

Suponiendo que la supervivencia de la censura residual R^* no está relacionada con la distribución de (Z^0, T^0) , se llega al resultado anunciado. $\square$

3. ESTUDIO DE SIMULACIÓN

En esta sección se estudia el comportamiento del estimador EMV mediante simulaciones. A efectos comparativos se incluyen en el estudio de simulación otros dos estimadores existentes para la supervivencia libre de evento bajo diseño de corte transversal (de Uña-Álvarez, 2017). El primero de ellos se puede ver como un caso particular de la integral límite-producto investigada por Sánchez-Sellero et al. (2005). Este estimador, que llamaremos estimador límite-producto integral (ELPI), se define como la primera marginal de

$$\hat{F}_{Z^0 T^0}^{ELPI}(z, t) = \sum_{i=1}^n W_i I(Z_i \leq z, T_i \leq t),$$

donde W_i es el salto en tiempo T_i del estimador límite-producto de Tsai et al. (1987) de la supervivencia total. Es decir, $\hat{F}_{Z^0}^{ELPI}(z) = \hat{F}_{Z^0 T^0}^{ELPI}(z, \infty)$. Un inconveniente de este estimador es que no asigna masa a los Z_i no censurados con T_i censurados, resultando ineficiente. Además, la consistencia está garantizada únicamente cuando el soporte de la censura contiene al soporte de los tiempos de vida. En caso contrario, $\hat{F}_{Z^0}^{ELPI}(z)$ puede presentar un sesgo sistemático que se hace más notorio en la cola derecha. En concreto, si b_T es el límite superior del soporte de T , entonces se verifica:

$$\hat{F}_{Z^0}^{ELPI}(z) \rightarrow P(Z^0 \leq z, T^0 \leq b_T).$$

El límite no coincide en general con la cantidad a estimar. No obstante, el ELPI es recomendable cuando no hay censura, ya que los dos inconvenientes que hemos comentado desaparecen en tal caso. De hecho, puede probarse que en el caso no censurado el ELPI puede introducirse como un

estimador de máxima versosimilitud.

El segundo de los estimadores es el estimador basado en el factor de reducción (EFR), que se define como una porción tiempo variante del estimador de Tsai et al. (1987):

$$\hat{S}_{Z^0}^{EFR}(z) = \hat{S}_{T^0}(z) \frac{\hat{K}_Z(z)}{\hat{K}_T(z)},$$

donde $\hat{S}_{T^0}(z)$ es el estimador de Tsai et al. (1987), $\hat{K}_Z(z) = \frac{1}{n} \sum_{i=1}^n I(L_i \leq z \leq Z_i)$ y $\hat{K}_T(z) = \frac{1}{n} \sum_{i=1}^n I(L_i \leq z \leq T_i)$. Nótese que el factor $\hat{K}_Z(z)/\hat{K}_T(z)$ es la proporción de casos con $Z_i \geq z$ entre los que satisfacen $L_i \leq z \leq T_i$. El EFR es consistente, pero presenta mayor varianza que el ELPI en escenarios con poca censura (de Uña-Álvarez, 2017). Una peculiaridad del EFR es que no salta únicamente en los Z_i , sino también en los T_i y en los L_i . Además, $\hat{S}_{Z^0}^{EFR}(z)$ no es monótono decreciente; este problema puede solucionarse considerando en la práctica su versión monotonizada $\hat{S}_{Z^0}^{EFR,m}(z) = \inf_{v \leq z} \hat{S}_{Z^0}^{EFR}(v)$, si bien ésta no ofrece mejoras notables en cuanto a varianza.

Los modelos simulados son los siguientes. Para el par (Z^0, T^0) consideramos $Z^0 \sim \text{Exp}(3)$ y $T^0 = Z^0 + \gamma V^0$, donde $\gamma \sim \text{Ber}(0.7)$ y $V^0 \sim \text{Exp}(1)$ son generadas independientemente de Z^0 . La variable de truncamiento, independiente de (Z^0, T^0) , sigue un modelo uniforme, $L \sim U(0, 2)$; mientras que la variable de censura C se genera de diferentes formas para simular los dos modelos (A1)–(A2) y (B1)–(B3) considerados en la Sección 2:

- Modelo A: $C = \eta C_1 + (1 - \eta)C_2$, donde $\eta \sim \text{Ber}(0.8)$, $P(C_1 = 2) = 1$ y $C_2 \sim U(0, 2 - L)$.
- Modelo B: $C = L + \tau$, con $\tau = 0.5$ (Modelo Ba), $\tau = 1$ (Modelo Bb), y $\tau = \infty$ (Modelo Bc, caso no censurado).

Nótese que en el modelo B el tiempo residual de censura $C - L$ es constante y, por tanto, verifica la condición (B3). Por otra parte, para los modelos A, Ba y Bb el soporte de C está estrictamente contenido en el de T^0 . Para simular una situación censurada donde esto no ocurre, hemos considerado un escenario más para el modelo B en el caso $\tau = 0.5$, rectificando el soporte de las variables de interés. En concreto, en nuestro modelo Bd la variable Z^0 se condiciona a $Z^0 \leq 0.65$ y V^0 se condiciona a $V^0 \leq 1.85$; esto implica que el límite superior del soporte de T^0 es 2.5, el cual coincide con el de $C = L + 0.5$. El sesgo del estimador ELPI es nulo en los escenarios Bc y Bd.

A efectos ilustrativos, en la Figura 3 se muestran las supervivencias libres de enfermedad estimadas por los tres estimadores (ELPI, EFR, EMV) para una muestra simulada del modelo Bb, con tamaño muestral $n = 60$. Se puede ver claramente que EMV se ajusta bien la curva teórica. Además es fácil apreciar que ELPI presenta menos puntos de salto (pues asigna masa nula a aquellos Z_i no censurados con T_i censurados) y que EFR, además de no ser monótono decreciente, salta en más puntos (en Z_i , T_i y L_i).

$n =$	ELPI			EFR			EMV		
	100	250	500	100	250	500	100	250	500
Modelo A	2	1	0.3	17.5	16.8	19.2	80.5	82.2	80.5
Modelo Ba	12.4	6.3	1.9	22	40.2	58.8	65.6	53.5	31.3
Modelo Bb	12.9	7.2	4.8	12.5	20.8	28.9	74.6	72	66.3
Modelo Bc	38.2	43.4	43.2	16.7	15.6	19.8	45.1	41	37
Modelo Bd	36.4	45.4	45.5	9.6	12	17.6	54	42.6	36.9

Tabla 1: Porcentaje de veces que funciona mejor cada estimador en términos del error integrado.

Para cada uno de los cinco modelos se han arrojado 1000 trials de Monte Carlo con tamaños muestrales $n = 100, 250, 500$. En la Tabla 1 se muestra el porcentaje de veces que funciona mejor cada uno de los estimadores en términos del error integrado, y en la Tabla 2 se muestra la mediana

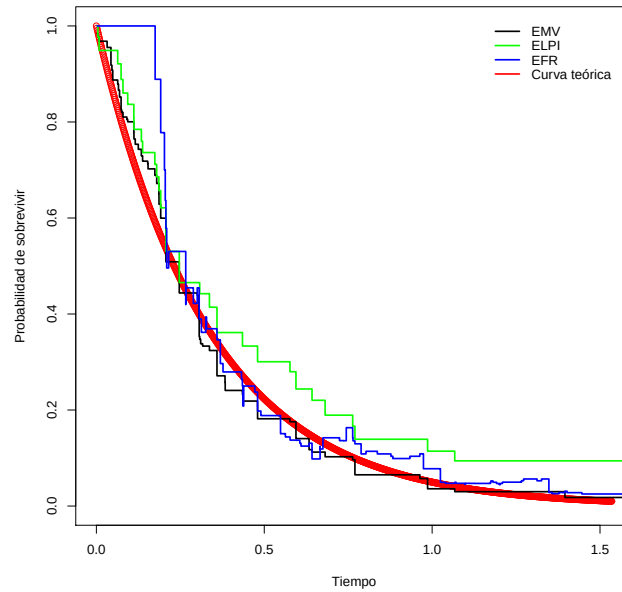


Figura 3: Supervivencia libre de enfermedad teórica y estimada mediante EMV, ELPI y EFR a partir de una muestra simulada del modelo Bb con $n = 60$ datos.

$n =$	ELPI			EFR			EMV		
	100	250	500	100	250	500	100	250	500
Modelo A	14.975	13.56	13.405	2.8765	1.2595	0.6126	1.2947	0.6279	0.3127
Modelo Ba	9.0882	6.8108	6.4284	4.0832	1.6338	0.8617	2.1300	1.2675	1.037
Modelo Bb	4.6878	3.3570	2.776	2.6155	1.1277	0.5558	1.0812	0.5386	0.3192
Modelo Bc	1.0432	0.4847	0.2157	2.237	1.0079	0.4006	1.0432	0.4847	0.2157
Modelo Bd	3.7562	1.5980	0.8773	6.5640	2.9945	1.3832	2.8914	1.4946	0.9225

Tabla 2: Mediana del ISE $\times 10^{-3}$.

del error integrado en cada escenario de simulación. Sean $x_1, \dots, x_r$, con $r = 100$, un conjunto de puntos equiespaciados en el soporte de T . Entonces, el error integrado se calcula como sigue: $ISE = \frac{1}{100} \sum_{i=1}^r (\hat{S}_{Z^0}(x_i) - S_{Z^0}(x_i))^2$. En las Figuras 4, 5 y 6 se muestran los diagramas de cajas correspondientes. Los resultados obtenidos muestran que:

- (a) la mediana del ISE del EMV decrece a medida que aumenta el tamaño muestral, sugiriendo la consistencia del nuevo estimador;
- (b) el EMV es el mejor estimador en la mayor parte de los escenarios simulados; excepciones a esto son el modelo Ba (donde EFR es mejor para $n = 500$) y el modelo Bd (donde ELPI es mejor también para $n = 500$); simulaciones adicionales (no mostradas) sugieren que el algoritmo utilizado para computar el EMV podría tener cierto sesgo en tales casos;
- (c) en el escenario no censurado (modelo Bc) el comportamiento de los estimadores EMV y ELPI es similar (nótese que, de hecho, para el caso no censurado el algoritmo iterativo computa el ELPI).

Un inconveniente del algoritmo introducido en la Sección 2 para computar el EMV es que no siempre converge. Es decir, en un número de casos el error excede el umbral prefijado (10^{-5}) tras

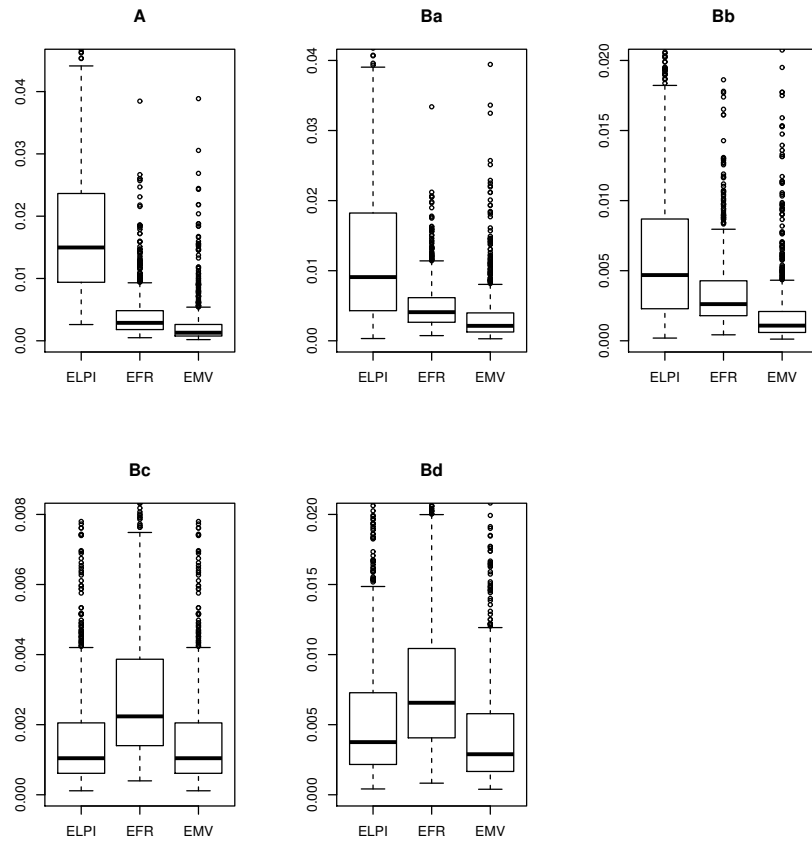


Figura 4: Diagramas de cajas del ISE para $n = 100$.

150 iteraciones. Esto ha sido especialmente frecuente en el modelo Ba, con un 51 % de no convergencias para $n = 100$, un 78 % para $n = 250$, y un 89 % para $n = 500$. El modelo Ba tiene una censura relativamente pesada (70 % sobre T^0 , 14 % sobre Z^0), lo cual podría tener influencia en el comportamiento del algoritmo. Para el resto de modelos, con censura más ligera, el algoritmo convergió en la mayoría de los casos. Procede no obstante una revisión más detallada del algoritmo que pueda aportar mejoras a este respecto. Por otra parte, el diseño del algoritmo en la Sección 2 conlleva la posibilidad de que aparezcan denominadores nulos o pesos negativos; para tales casos hemos implementado correcciones (conservar la solución de la iteración anterior o acotar inferiormente los pesos por cero, respectivamente). Otras correcciones que puedan ser más eficientes están actualmente bajo estudio.

4. APLICACIÓN AL TIEMPO LIBRE DE INFECCIÓN EN UCIs

En esta Sección analizamos a efectos ilustrativos los datos de las UCIs introducidos en la Sección 1. Recuerdese que el evento intermedio de interés es BSI y que los posibles estados en los que se puede encontrar un paciente son: hospitalizado sin BSI, hospitalizado con BSI, y salida del hospital (ya sea por alta o por muerte en la UCI). El seguimiento de los pacientes se limita a 30 días desde el ingreso en la UCI, por lo que son censurados en dicho tiempo. Se tiene la información de 134 pacientes de los cuales 29 murieron en la UCI, 81 fueron dados de alta vivos en la UCI y 24 fueron censurados (20 de ellos en 30 días, los 4 restantes durante el seguimiento). Para los 41 individuos que sufrieron BSI se conoce el tiempo hasta tener dicha infección. Estos datos de UCIs se adaptan al modelo (A1)–(A2) descrito en la Sección 2.

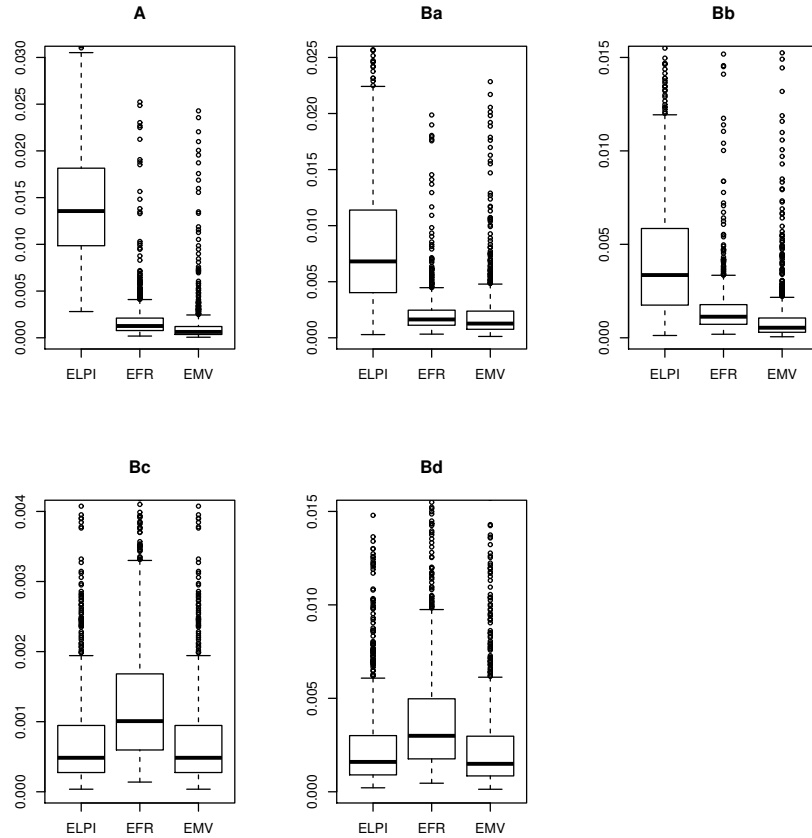


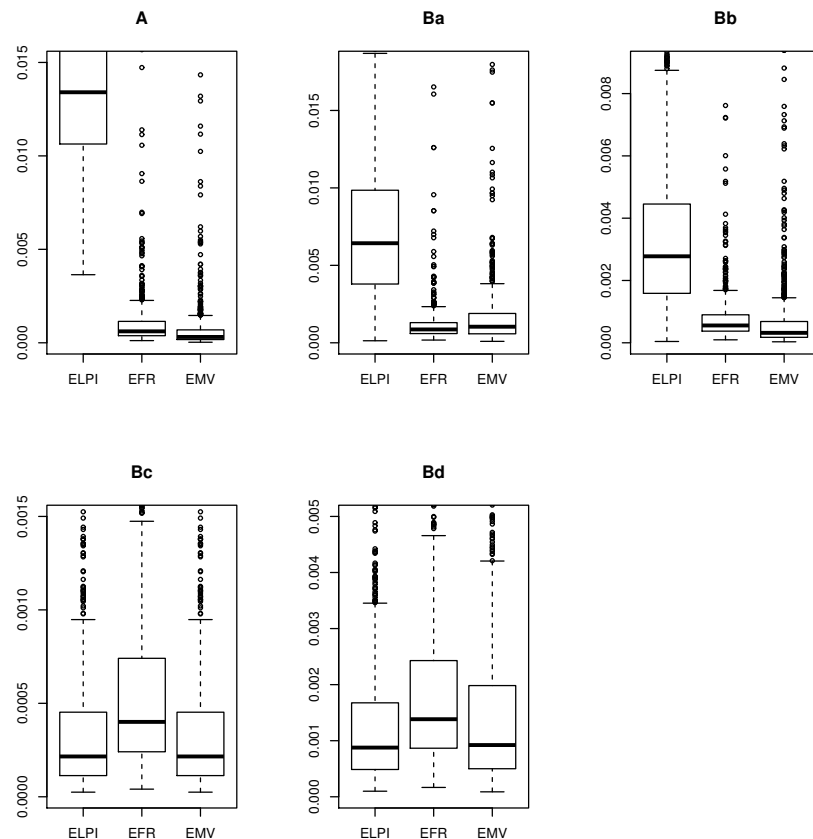
Figura 5: Diagramas de cajas del ISE para $n = 250$.

En la Figura 7, izquierda se muestra la estimación de la supervivencia en la UCI libre de infección obtenida mediante los estimadores EMV, ELPI y EFR. Las tres curvas arrojan resultados similares. En la Figura 7, derecha se muestra de nuevo el estimador EMV, acompañado ahora por los límites de confianza puntuales al 95 % basados en el bootstrap simple y en el método percentil. Dichos límites son de la forma $(q_{0.025}^{boot}, q_{0.975}^{boot})$, siendo q_{α}^{boot} el cuantil α de los valores del EMV a lo largo de las remuestras bootstrap (se consideraron 500 remuestras). En esta última figura se incluye también el estimador Kaplan-Meier ordinario, que ignora el truncamiento por la izquierda; la gráfica permite observar cómo el método Kaplan-Meier usual sobreestima largamente la supervivencia libre de infección.

5. COMENTARIOS FINALES Y CONCLUSIONES

En este trabajo se ha introducido un nuevo estimador para la supervivencia libre de enfermedad con datos de corte transversal. El estimador propuesto se ha construido en base a un criterio de máxima verosimilitud no paramétrica. Como el estimador no tiene forma explícita, se ha propuesto un algoritmo iterativo para su cómputo en la práctica. La consistencia del estimador se ha investigado mediante estudios de simulación intensivos. Estos estudios han mostrado que el estimador propuesto es más preciso que otros estimadores existentes. A modo ilustrativo, el EMV se ha utilizado para estimar la supervivencia libre de infección en UCIs.

Una cuestión interesante es si, bajo muestreo de corte transversal, la verosimilitud condicional de los (Z_i, δ_i) dados los L_i , $1 \leq i \leq n$, contiene toda la información relevante sobre la distribución de Z^0 , o si el EMV puede mejorarse mediante el uso de los (T_i, Δ_i) . Otro tema a investigar es la

Figura 6: Diagramas de cajas del ISE para $n = 500$.

posible mejora del algoritmo iterativo propuesto para el cómputo del EMV el cual, como se ha dicho, puede presentar problemas de convergencia. Relacionado con esto, sería deseable disponer de una fórmula explícita para el EMV que hiciera innecesario un proceso de optimización numérica; los esfuerzos realizados hasta ahora han sido infructuosos.

También como trabajo futuro, sería interesante analizar las propiedades del EMV desde un punto de vista teórico, así como la consistencia del bootstrap simple en la aproximación de errores estándar y la construcción de intervalos de confianza. Otro objetivo es el empaquetado del código R utilizado para la implementación del estimador, de tal manera que pueda ser utilizado de manera sencilla por la comunidad científica.

AGRADECIMIENTOS

Este trabajo fue realizado durante el contrato de la primera autora como Técnico de Apoyo a la Investigación del CINBIO. Los autores agradecen el soporte económico de la Xunta de Galicia (acreditación del CINBIO como centro singular de investigación de Galicia, 2016-2019) y de la Unión Europea (European Regional Development Fund - ERDF). El segundo autor agradece la financiación del proyecto MTM2017-89422-P (MINECO/AEI/FEDER,UE).

REFERENCIAS

- Buyse, M., Burzykowski, T., Carroll, K. et al. (2007) Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(33), 5218-5224.
- de Uña-Álvarez, J. (2017) Nonparametric Estimation of an Event-Free Survival Distribution Under Cross-Sectional Sampling. Ferger D., González Manteiga W., Schmidt T., Wang JL. (eds) *From Statistics to Mathematical Finance*. Springer.

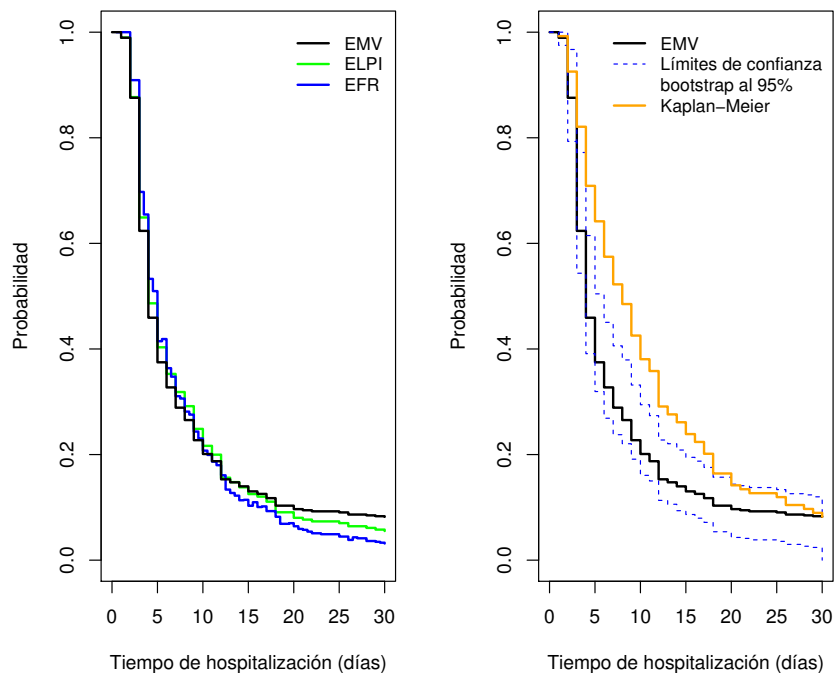


Figura 7: Supervivencia libre de infección en UCIs.

- de Uña-Álvarez, J. y Mandel, M. (2018) Nonparametric estimation of transition probabilities for a general progressive multi-state model under cross-sectional sampling. *Biometrics*, 74(4), 1203-1212.
- Fluss, R., Mandel, M., Freedman, L. S. et al. (2013) Correction of sampling bias in a cross-sectional study of post-surgical complications. *Statistics in Medicine*, 32, 2467-2478.
- Gijbels, I. y Wang, J. (1993) Strong representations of the survival function for truncated and censored data with applications, *Journal of Multivariate Analysis*, 47, 210-229.
- Kaplan, E.L. y Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.
- Korn, R. L. y Crowley, J. J. (2013) Overview: Progression-free survival as an endpoint in clinical trials with solid tumors. *Clinical Cancer Research*, 19, 2607-2612.
- Sánchez-Sellero, C., González-Manteiga, W. y Van Keilegom, I. (2005) Uniform representation of product-limit integrals with applications. *Scandinavian Journal of Statistics*, 32, 563-581.
- Tsai, W. Y., Jewell, N. P. y Wang, M. C. (1987) A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74, 883-886.
- Vakulenko-Lagun, B. y Mandel, M. (2016). Comparing estimation approaches for the illness-death model under left truncation and right censoring. *Statistics in Medicine*, 35, 1533-1548.
- Vakulenko-Lagun, B., Mandel, M. y Goldberg, Y. (2017). Nonparametric estimation in the illness-death model using prevalent data. *Lifetime Data Analysis*, 23, 25-56.
- Wang, M.-C. (1991) Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86, 130-143.
- Zhou, Y. y Yip, P. (1999) A strong representation of the product-limit estimator for left truncated and right censored data, *Journal of Multivariate Analysis*, 69, 261-280.

Pósters

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

ESTADÍSTICA EN LA PLANIFICACIÓN ESTRATÉGICA DE ABANCA

Belén M Fernández de Castro¹, Teresa Veiga Rodríguez¹, Alejandro Figueroa Silva¹ y Sandra Amado Souto¹

¹ABANCA

RESUMEN

La Dirección General de Planificación Estratégica y PMO de ABANCA está formada por un equipo multidisciplinar que desempeña muy diversas tareas de relevancia para la Entidad. El área de Planificación y Estudios se encarga, entre otras, de hacer un seguimiento de la evolución macroeconómica, el desarrollo de los Planes Estratégicos, la presupuestación anual, dar respuesta a los distintos requerimientos del supervisor (Stress Test, ICAAP,...) o dar soporte a los órganos de decisión del Banco. Además, desde el área de Gobierno del Dato se vela por la calidad de la información que se maneja en la entidad.

El presente trabajo recoge una visión general de cómo la estadística puede dar soporte a diversas tareas en el marco de la Planificación Estratégica de una entidad financiera.

Palabras y frases clave: Series temporales, modelos macroeconómicos, R y calidad del dato

1. INTRODUCCIÓN

Con el fin de dar una visión global de las aplicaciones de la estadística en el departamento de Planificación Estratégica, se describirán los datos con los que habitualmente se trabaja, así como la importancia y el esfuerzo que se está haciendo dentro de la entidad para garantizar la calidad de la información utilizada.

A continuación, se presentan algunas de las técnicas estadísticas que se emplean para el estudio de series temporales, bien con el fin de cumplir los requisitos del regulador o como parte del seguimiento macroeconómico. Se muestran además las herramientas empleadas para registrar, compartir y facilitar los resultados de los estudios.

Por último, se presentan tres casos prácticos que ponen de manifiesto las técnicas detalladas anteriormente. La mayor parte del trabajo se desarrolla en el software estadístico R, por lo que se hará alusión a algunos de los paquetes empleados.

2. DATOS UTILIZADOS

Entre los datos con los que se trabaja dentro de ABANCA se puede distinguir entre:

- Datos internos. En ABANCA, como entidad financiera, se están generando datos de manera continua (por ejemplo, cada vez que un cliente realiza una operación). Estos datos se emplean con muy diversas finalidades, tanto de reporte y seguimiento como en modelización en el ámbito de riesgos (cálculo de límites de crédito) o en el ámbito de marketing (modelos de propensión o recomendación)
- Datos externos. Como parte del seguimiento macroeconómico y de las modelizaciones, se emplean datos facilitados por distintos organismos públicos como son el Banco de España, INE, IGE o Eurostat. Mediante una función programada en R, se ha automatizado la descarga de toda esta información.

En un entorno en el que se dispone de gran cantidad de datos de muy diversa naturaleza y que abarcan muy distintos ámbitos, resulta de vital importancia garantizar la calidad de la información que se pone a disposición de los usuarios, tanto para su reporte como para su utilización en modelos estadísticos en los ámbitos de marketing, riesgos y planificación. En este sentido, la Entidad está desarrollando un Sistema de Control de Calidad de Datos que, mediante controles automáticos implementados sobre la base de datos, garantiza unos estándares mínimos de calidad en 6 dimensiones: completitud, unicidad, disponibilidad, validez, consistencia e integridad.

3. TRATAMIENTO, MODELIZACIÓN Y DIFUSIÓN DE LOS RESULTADOS

En el ámbito de trabajo de las series temporales, para evitar sacar conclusiones erróneas, es necesario corregir los datos de atípicos, efectos de calendario y estacionalidad.

- En las variables de carácter interno, la corrección de los datos atípicos está motivada por acontecimientos puntuales que no tienen que ver con el ciclo económico (fusiones y ventas de carteras, por ejemplo)
- Para no llegar a resultados engañosos al comparar distintos periodos es necesario tener en cuenta la estacionalidad. Variables como por ejemplo el gasto en tarjetas de débito tiene una fuerte componente estacional
- Los efectos de calendario son movimientos que están relacionados con la composición del calendario, como la Semana Santa, el número de días laborables y el año bisiesto. Variables como la facturación de los TPVs de ABANCA se ven afectadas por el efecto de la Semana Santa, fiesta móvil del calendario.

Una vez depuradas las series, es posible analizarlas, obtener conclusiones y, en la mayoría de los casos, será de interés obtener predicciones a futuro. Dentro de las modelizaciones que se llevan a cabo con este fin, se emplean principalmente modelos Box-Jenkins, regresiones dinámicas, modelos no lineales, modelos multivariantes y componentes principales dinámicas. Toda esta metodología se pone en práctica en el software estadístico R, combinando funciones de desarrollo propio y los paquetes TSA, gdp y GAM, entre otros.

A la hora de elegir un modelo, se prioriza:

- Validez estadística
- Medida del grado de ajuste
- Buen nivel predictivo: ejercicio de backtesting
- Buena respuesta a los análisis de sensibilidad

Una vez obtenidos los resultados es importante tanto facilitar su difusión y utilización por otros usuarios internos, como la generación de documentación que permita tener una traza y justificación de los resultados utilizados en los reportes. Para ello, se emplean dos vías:

- Generación automática de documentos con R Markdown
- Creación de aplicaciones con R Shiny, que permite a los usuarios no conocedores de R la visualización mediante tablas o gráficas dinámicas de los datos, análisis de los modelos y estudios de sensibilidad de las predicciones obtenidas con los modelos, entre otras funcionalidades.

4. CASOS PRÁCTICOS

A continuación, se detallan algunos de los casos prácticos de modelizaciones y presentación de resultados en los que se ha trabajado dentro del Departamento.

- IFRS9. Generación de escenarios base, optimista y adverso a 3 años con asignación de probabilidad. Se han desarrollado modelos estadísticos que permiten generar un cuadro de indicadores macroeconómicos, con previsiones a 3 años, que sirven de soporte a toda la entidad. Estos modelos macroeconómicos se han desarrollado con el software estadístico R. Además, se ha implementado una aplicación que permite a los usuarios interactuar de manera sencilla permitiendo, por ejemplo, actualizar los datos, revisar las previsiones y los modelos estimados y obtener un cuadro de previsiones completo

- Presupuestación. Mediante modelizaciones se vinculan ciertas variables de negocio al escenario macroeconómico establecido para proyectar su evolución y plantear unos objetivos en términos de ganancia de cuota de mercado
- Construcción de indicadores de seguimiento mensual de las distintas componentes del PIB. A partir de variables internas de la Entidad se han construido dos indicadores para el consumo privado de las familias gallegas y la inversión de las empresas gallegas. La elevada cuota de mercado de ABANCA en Galicia hace que sus clientes sean una muestra representativa de la comunidad, lo que permite y motiva la generación de estos modelos.

REFERENCIAS

- Daniel Peña, Ezequiel Smucler and Victor Yohai (2018). `gdpc`: Generalized Dynamic Principal Components. R package version 1.1.0. <https://CRAN.R-project.org/package=gdpc>
- JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng and Winston Chang (2018). `rmarkdown`: Dynamic Documents for R. R package version 1.10. <https://CRAN.R-project.org/package=rmarkdown>.
- Kung-Sik Chan and Brian Ripley (2012). `TSA`: Time Series Analysis. R package version 1.01. <https://CRAN.R-project.org/package=TSA>.
- Ruey S. Tsay (2015). `MTS`: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 0.33. <https://CRAN.R-project.org/package=MTS>.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Sandra Amado Souto (2018). Previsión de variables de negocio y generación de escenarios macroeconómicos con modelo estadísticos. Trabajo Fin de Máster (USC).
- Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). `shiny`: Web Application Framework for R. R package version 1.1.0. <https://CRAN.R-project.org/package=shiny>.

ESTUDO DA DIVERSIDADE XENÉTICA EN DÚAS POBOACIÓNS DE *Zostera noltei*

Álvarez, I.¹, Hermida, J.¹ e López, E.¹

¹ IES A Sangriña. Paseo de Portugal, 21. 36.780 A Guarda. Pontevedra.

RESUMO

Neste traballo preséntanse os resultados preliminares da caracterización da diversidade xenética de dúas poboacións de *Zostera noltei*. *Z. noltei* é unha planta fanerógama marina que forma extensas pradeiras no inter e submareal de moitas zonas estuáricas de Galicia. O forte declive das súas poboacións pon en risco non só un importante ecosistema senon tamén a viabilidade da propia especie. A diminución do tamaño poboacional ten consecuencias na perda de diversidade xenética e no illamento reproductivo entre subpoboacións. O estudo de marcadores xenéticos AFLPs ofrece certas vantaxes para o estudo da diversidade xenética de especies sobre as que non é necesario coñecer a súa secuencia de ADN. Neste traballo estudáronse 87 AFLPs para caracterizar a diversidade xenética de dúas poboacións de *Z. noltei*, unha ubicada na Ramallosa (Nigrán) e outra en Vilaboa. Os resultados indican elevados graos de diversidade xenética e a ausencia de diverxencia xenética entre ambas poboacións.

Palabras e frases chave: estimadores de diversidade xénica, AFLPs, estrutura poboacional.

1. INTRODUCCIÓN

Os proxectos de investigación constitúen unha estratexia adecuada na aprendizaxe de contidos no eido das ciencias. O IES A Sangriña desenvolve por primeira vez o programa STEMbach, en colaboración coa Universidade de Vigo. Alumnado de 1º de bacharelato deseñan e desenvolven proxectos de investigación orixinais baixo o obxectivo xeral de avaliar a calidade ambiental dunha zona costeira fortemente antropizada. Os 10 proxectos que se están desenvolvendo teñen por finalidade avaliar tres aspectos fundamentais, a) analizar os datos climáticos (temperatura e pluviosidade) recollidos no período 2003-2018 pola estación meteorolóxica do Rosal, para detectar posibles evidencias de cambio climático, b) avaliar o efecto nos seres vivos de axentes contaminantes tales como TBT, Hidrocarburos Aromáticos Policíclicos (HAPs) e metais pesados (Zn) en interacción con microplásticos mediante bioensaos co bivalvo *Corbicula fluminea* e co ourizo de mar *Paracentrotus lividus* e da alteración granulométrica dos sedimentos e con materia orgánica de orixe antropoxénica, c) avaliar os efectos da diminución das poboacións naturais a través de análises de diversidade xenética con AFLPs en *Z. noltei* e da estrutura xenética ADN mitocondrial en *Charadrius alexandrinus*.

Neste traballo preséntanse os resultados preliminares da caracterización da diversidade xenética de dúas poboacións de *Z. noltei*.

As pradeiras submarinas de *Z. noltei* atópanse amplamente distribuídas na sección inter e submareal das costas de Europa. Estas pradeiras submarinas contribúen á creación de importantes hábitats. Nalgunhas áreas estas poboacións experimentaron un significativo declive que conleva, entre outras consecuencias, a perda de diversidade xenética que pode conducilas á súa extinción. O estudo da diversidade xenética destas poboacións contribúe a coñecer o grao de diferenciación xenética entre poboacións, os mecanismos reprodutores así como as estratexias de protección e conservación.

O uso de AFLPs como marcadores xenéticos ofrece algunhas vantaxes fronte ós microsátélites, marcadores moleculares altamente polimórficos e amplamente usados neste tipo de estudos. Non é preciso coñecer a secuencia de ADN da especie obxecto de estudo e constitúe unha metodoloxía doada, rápida e de baixo custo. Sen embargo, os AFLPs presentan principalmente dous problemas para a súa análise. Por un lado, fragmentos de ADN distintos poden ser asignados ó mesmo locus, fenómeno coñecido como homoplasia, e por outro, os individuos homocigotos para un alelo amplificable non poden

diferenciarse dos individuos heterozigotos para ambos alelos amplificable e non-amplificable. Aínda que estes inconvenientes non impiden a estimación de frecuencias alélicas para a análise xenético-poboacional, reducen a exactitude de tales estimacións con respecto a outros marcadores codominantes (Lynch e Milligan, 1994). Neste traballo aplicanse estimadores desenvolvidos para este tipo de marcadores con dominancia. Todos estes estimadores teñen en conta algún sesgo debido ó pequeno tamaño poboacional (Lynch e Milligan, 1994).

2. MATERIAL Y MÉTODOS

Recolléronse 5 plantas de dúas poboacións illadas, Ramallosa (Nigrán) e Vilaboa (ambas na provincia de Pontevedra, Galicia), distantes entre si 30 km, e analizáronse 87 AFLPs. Na recolección de individuos tívose en conta que non procedesen da mesma mata a fin de evitar a presenza de individuos orixinados por reprodución asexual, xenéticamente idénticos. O xenotipado dos individuos realizouse co software *peak scanner software v1.0* e a análise de datos fíxose co software R.

A análise de estrutura poboacional levouse a cabo según se describe en Lynch e Milligan (1994).

Aplicouse o estimador de frecuencias alélicas baixo a asunción de equilibrio Hardy-Weinberg (H-W):

$$\hat{q} = \hat{x}^{1/2} \left(1 - \frac{Var(\hat{x})}{8\hat{x}^2} \right)^{-1}, \text{ onde,}$$

$\hat{x}$, é a proporción dos N individuos da mostra que non presentan o marcador (alelo nulo) e $Var(\hat{x}) = \hat{x}(1 - \hat{x})/N$.

A heterozigosidade esperada asumindo equilibrio H-W foi calculada mediante o estimador:

$$\hat{H}_j(i) = 2\hat{q}_j(i)[1 - \hat{q}_j(i)] + 2Var[\hat{q}_j(i)]$$

cuxa varianza é aproximadamente $Var[\hat{H}_j(i)] = 4[1 - 2\hat{q}_j(i)]^2 Var[\hat{q}_j(i)]$.

3. RESULTADOS

- Estimación de frecuencias xénicas.

A estimación de frecuencias xénicas pode presentar un sesgo cando o tamaño da mostra é pequeno ou cando as frecuencias xénicas son extremas (frecuencia do alelo nulo < 0,1). As frecuencias do alelo nulo na poboación de Ramallosa oscilaron entre 0,31 e 0,92 e na poboación de Vilaboa entre 0,33 e 0,92, polo que en ningún caso se tiveron en conta loci con frecuencias xénicas inferiores a 0,1.

- Diversidade xenética.

A diversidade xenética dunha poboación pode medirse en función do valor de heterozigosidade que presenta.

Para loci dialélicos, a heterozigosidade esperada oscila entre 0, para frecuencias alélicas extremas e 0,5 para frecuencias alélicas intermedias. Como se pode observar na táboa 1, a heterozigosidade esperada en ambas poboacións é elevada.

	N	H_j	S.E.
Ramallosa	5	0.43376	0.01190
Vilaboa	5	0.42976	0.01225

Táboa 1: Heterozigosidade esperada (H_j) baixo equilibrio H-W.

Considerando as dúas poboacións conxuntamente, a diversidade xenética promedio (táboa 2) foi estimada según a expresión

$$\hat{H}_w = \frac{1}{n} \sum_{j=1}^n \hat{H}_j$$

n	Hw	S.E.(Hw)
2	0.4318	0.00200

Táboa 2: Diversidade xenética promedio (Hw) das poboacións de Ramallosa e Vilaboa.

- Estrutura xenética das poboacións.

Cando se dispoñen de datos de máis dunha poboación, é interesante avaliar o grao no que os compoñentes de diversidade xénica entre (Hb) e dentro (Hw) das poboacións contribúen á diversidade xénica total (Ht). Os resultados poden verse na táboa 3.

n	Ht	Hw	Hb	Fst
2	0.4234	0.4318	-0.0083	-0.0197
S.E.		0.002000	0.000000	0.004720

Táboa 3: Descomposición dos compoñentes que contribúen á diversidade xénica.

Como xa se dixo con anterioridade, a diversidade xenética das poboacións analizadas é alta á vez que moi semellante en ambas. A contribución da diversidade xenética entre as poboacións á diversidade total é moi baixa (Hb=-0,0083). O valor Fst mide o grao de subdivisión dunha poboación. O estimador definido por Wright (1951) ofrece unha estima da redución da heterocigosidade nunha subpoboación debido ós cruzamentos non ó azar, con respecto á poboación total. O valor Fst oscila entre 0 cando non existe diferenciación entre subpoboacións e as frecuencias xénicas son idénticas, e 1 cando se produciu a fixación de alelos alternos en diferentes subpoboacións. Nalgunhas ocasións Fst pode ser negativo a causa das características das mostras. O valor de Fst calculado indica a ausencia de diferenciación xenética entre as poboacións analizadas.

- Distancia xenética.

Nei (1972) introduce a medida de distancia xenética entre poboacións que ofrece unha estima do número medio de mutacións xénicas que separan ambas poboacións. A partir dos datos obtidos a distancia xenética de Nei estimada a partir do estimador de Lynch e Milligan (1994) é 0,0000. Non se observa diverxencia xenética entre ambas poboacións.

4. CONCLUSIÓNS

1.- A diversidade xenética das poboacións de *Z. noltei* da Ramallosa e Vilaboa avaliada mediante marcadores xenéticos AFLPs é elevada.

2.- Non existe diverxencia xenética entre as poboacións de *Z. noltei* de Ramallosa e Vilaboa.

REFERENCIAS

- Lynch, M. e Milligan, G. (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology*, 3, 91-99.
- Nei, M. (1972) Genetic distance between populations. *American Naturalist*, 106, 283-292.
- Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics*, 15, 323-354.

Agradecementos:

Este traballo foi financiado polo programa STEMBach da Consellería de Educación, Universidade e Formación Profesional da Xunta de Galicia. O noso agradecemento ós profesores da Universidade de Vigo, Emilio Fernández, codirector do proxecto, Paloma Morán e Juan Galindo do Departamento de Xenética polo seu asesoramento e apoio técnico e a Gloria Fiestras do Departamento de Estatística e Investigación Operativa polos seus consellos.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

EVIDENCIAS LOCAIS DO CAMBIO CLIMÁTICO A PARTIR DO REXISTRO DE DATOS DAS ESTACIÓNS METEOROLÓXICAS DO ROSAL E DE OIA

Fernández, M.¹, López, E.¹, Hermida, J.¹ e Saavedra, A.²

¹ IES A Sangriña. Paseo de Portugal, 21. 36.780 A Guarda. Pontevedra.

² Departamento de Estatística e Investigación Operativa. Universidade de Vigo.

RESUMO

Neste traballo preséntanse os resultados preliminares da análise de series temporais dos datos de temperatura e pluviosidade rexistrados no período 2003-2018 na estación meteorolóxica do Rosal. Os datos foron analizados co software R.

A análise da serie temporal indica que se produciu un lixeiro incremento da temperatura media no Rosal. Este incremento non se produciu de xeito homoxéneo en todos os meses. Así mesmo, a temperatura máxima e mínima sufriron un incremento no mesmo período de tempo analizado. Con respecto ós datos rexistrados na estación meteorolóxica de Oia, produciuse un incremento da temperatura media mentres que a análise de tendencia estacional de Mann-Kendall indica que esta tendencia foi homoxénea en todos os meses. A temperatura máxima experimentou un leve incremento mentres que a temperatura mínima non experimentou cambios no período de tempo analizado. O feito de que a temperatura mínima non variara pode atribuírse á influencia oceánica nas características climáticas locais.

Non se atoparon evidencias significativas que indiquen que a precipitación media variase no rexistro do Rosal no período analizado. Sen embargo, o rexistro da estación meteorolóxica de Oia indica un importante incremento da precipitación. A análise de tendencia estacional de Mann-Kendall indica que este incremento foi homoxéneo en todos os meses excepto no mes de outubro.

Palabras e frases chave: Series de tempo, análise de tendencias, cambio climático.

1. INTRODUCCIÓN

Os proxectos de investigación constitúen unha estratexia adecuada na aprendizaxe de contidos no eido das ciencias. O IES A Sangriña desenvolve por primeira vez o programa STEMBach, en colaboración coa Universidade de Vigo. Alumnado de 1º de bacharelato deseñan e desenvolven proxectos de investigación orixinais baixo o obxectivo xeral de avaliar a calidade ambiental dunha zona costeira fortemente antropizada. Os 10 proxectos que se están desenvolvendo teñen por finalidade avaliar tres aspectos fundamentais, a) analizar os datos climáticos (temperatura e pluviosidade) recollidos no período 2003-2018 pola estación meteorolóxica do Rosal, para detectar posibles evidencias de cambio climático, b) avaliar o efecto nos seres vivos de axentes contaminantes tales como TBT, Hidrocarburos Aromáticos Policíclicos (HAPs) e metais pesados (Zn) en interacción con microplásticos mediante bioensaio co bivalvo *Corbicula fluminea* e co ourizo de mar *Paracentrotus lividus* e da alteración granulométrica dos sedimentos e con materia orgánica de orixe antropoxénica, c) avaliar os efectos da diminución das poboacións naturais a través de análises de diversidade xenética con AFLPs en *Zostera noltei* e da estrutura xenética ADN mitocondrial en *Charadrius alexandrinus*.

Neste traballo preséntanse os resultados preliminares da análise dos datos climáticos (temperatura e pluviosidade) correspondentes ao período 2003-2018 rexistrados na estación meteorolóxica do Rosal.

2. METODOLOXÍA

Unha serie de tempo é o resultado da observación sistemática dunha variable Z ao longo do tempo. Estas observacións teñen a restrición de que deben ser tomadas en intervalos regulares de tempo e, de forma xeral, asúmese que poden estar formadas por catro compoñentes:

$$Z_t = T_t + S_t + \varepsilon_t \quad (1)$$

Onde T_t é a tendencia e modeliza o comportamento a gran escala da serie, S_t é a compoñente estacional e recolle o comportamento periódico e ε_t denomínase erro e recolle a parte impredecible da serie de tempo.

Asumindo o modelo lineal indicado en (1), pódense estimar as compoñentes dunha serie de tempo co gallo de analizar:

1. Estudo de tendencia. Mediante a estimación de T_t analízase a posible existencia de evolucións ao longo do tempo.
2. Estudo da compoñente estacional. Trátase neste caso de chequear posibles derivas no comportamento periódico da serie.

Esta metodoloxía aplicouse ás series de tempo correspondentes aos datos de temperatura e pluviosidade rexistrados no período 2003-2018 na estación meteorolóxica do Rosal. Esta información é pública e pode descargarse libremente desde a web de MeteoGalicia: www.meteogalicia.gal

3. ESTUDO DA TEMPERATURA

Na Figura 1 móstrase a serie de tempo de temperatura media rexistrada na estación do Rosal e as estimacións das tres compoñentes teóricas T_t , S_t e ε_t . A representación gráfica das medicións de temperatura mostran un marcado comportamento periódico.

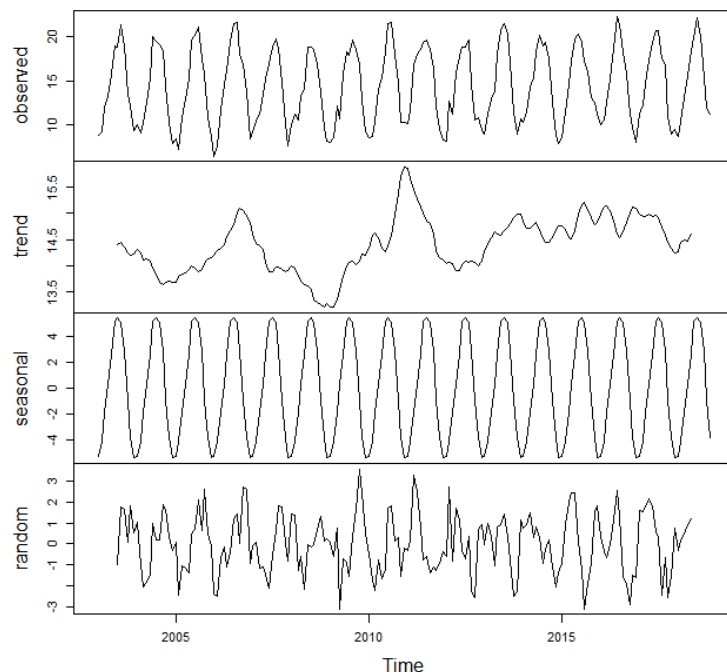


Figura 1. Temperatura media rexistrada na estación meteorolóxica do Rosal e estimación das súas compoñentes teóricas asumindo un modelo lineal.

Unha análise mediante regresión confirmou a existencia de tendencia positiva (coeficiente de determinación $R^2=0,2967$, $p < 0,05$) o que indica un aumento da temperatura ao longo dos anos.

Analizando a compoñente estacional mediante o test non paramétrico de Mann-Kendal, constatouse que, asumindo unha significación $\alpha=0,05$, nos meses de abril, maio, xuño, xullo e decembro pódese asumir un incremento da temperatura media nestes últimos anos. Por outra banda, no mes de outubro rexístrase un descenso da temperatura media.

Mes	Estatístico	P-valor	Mes	Estatístico	P-valor
Xaneiro	33	0,1492	Xullo	46	0,04234
Febreiro	42	0,0643	Agosto	-17	0,4708
Marzo	44	0,05239	Setembro	-40	0,07911
Abril	64	0,00456	Outubro	-46	0,04276
Maio	60	0,00777	Novembro	-41	0,07143
Xuño	46	0,04276	Decembro	53	0,00971

Táboa 1. Análise da compoñente estacional da temperatura media mediante o test non paramétrico de Mann-Kendal

Estes contrastes estatísticos repetíronse para o estudo das series de temperatura máxima e mínima recollidas na estación do Rosal obténdose resultados moi similares (coeficientes de determinación $R^2=0,3964$ e $R^2=0,1257$, respectivamente, $p < 0,05$).

A análise das series de tempo de temperaturas media, mínima e máxima rexistradas na estación meteorolóxica de Oia arroxaron resultados algo diferentes no sentido que, cunha significación $\alpha=0,05$, so se pode asumir un incremento das temperaturas media e máxima ao longo do tempo (coeficiente de determinación $R^2=0,02575$ e $R^2=0,1214$, respectivamente). A análise da compoñente estacional das tres series mediante o test non paramétrico de Mann-Kendal indica que esta tendencia foi homoxénea en todos os meses.

4. ESTUDO DA PLUVIOSIDADE

A representación gráfica da precipitación rexistrada na estación meteorolóxica do Rosal, Figura 2, pon de manifesto o peso da compoñente aleatoria ε_t , polo que todas as conclusións derivadas da análise da tendencia e da compoñente estacional deben tomarse con certa precaución.

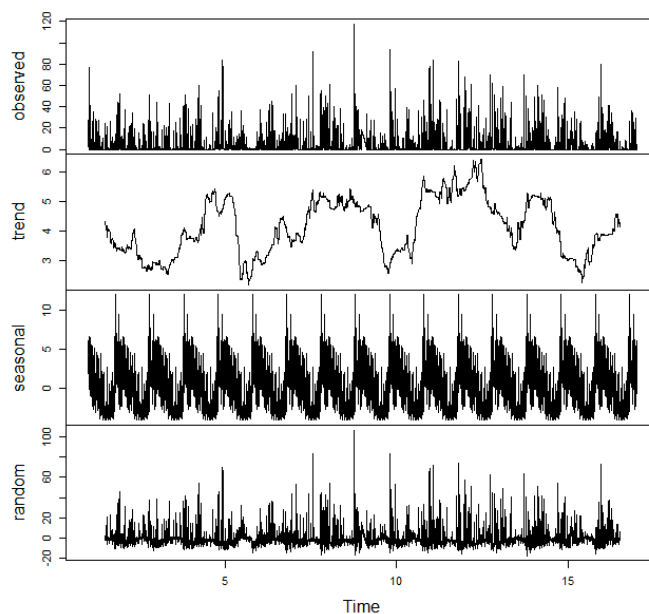


Figura 2. Precipitación rexistrada na estación meteorolóxica do Rosal e estimación das súas compoñentes teóricas asumindo un modelo lineal.

Os tests estatísticos non son concluíntes na determinación do incremento da precipitación media no rexistro do Rosal ao longo dos anos (coeficiente de determinación $R^2=0,07517$, $p < 0,05$). Sen embargo, o rexistro da estación meteorolóxica de Oia indica un importante incremento da precipitación (coeficiente de determinación $R^2=0,429$, $p < 0,05$). Tras aplicar o test non paramétrico de Mann-Kendall para estudar posibles fluctuacións da pluviosidade mes a mes detectouse que este incremento foi homoxéneo en todos os meses excepto no mes de outubro ($p < 0,05$).

5. CONCLUSIONES

Os resultados poñen de manifesto a influencia do relevo nas condicións climáticas locais. Mentres que Oia posúe unha meteoroloxía fortemente influenciada polo Océano Atlántico, con temperaturas máis suaves e maiores precipitacións por mor dos temporais, O Rosal desfruta dun clima máis “continental” con temperaturas máis elevadas e menores precipitacións. A Serra da Groba con altitudes de ata 600 m, supón unha barreira xeográfica causante do efecto fosen: as masas de aire oceánicas cargadas de humidade arrefriáanse e descargan en forma de precipitacións ó ascender a cotas de maior altitude, descendendo ó Val do Rosal con menor humidade e máis cálidas.

Para acadar os obxectivos iniciais do proxecto, isto é detectar evidencias de cambio climático, faise necesario ampliar o estudo a un maior número de estacións meteorolóxicas así como incrementar no posible o intervalo de tempo analizado. Aínda así, e tendo en conta esta limitación, a análise dos datos indica un lixeiro incremento da temperatura media, máxima e mínima no Rosal. Na estación meteorolóxica de Oia, produciuse un incremento da temperatura media. A temperatura máxima experimentou un leve incremento mentres que a temperatura mínima non experimentou cambios no período de tempo analizado. Con respecto ao estudo da pluviosidade, o rexistro da estación meteorolóxica de Oia indica un importante incremento da precipitación. A análise de tendencia estacional de Mann-Kendall indica que este incremento foi homoxéneo en todos os meses excepto no mes de outubro.

REFERENCIAS

- Cowpertwait, P.S.P. and Metcalfe, A.V. (2009). *Introductory Time Series with R*. Springer.
Peña, D. (2010). *Análisis de Series Temporales*. Alianza Editorial.

Agradecementos:

Este traballo foi financiado polo programa STEMBach da Consellería de Educación, Universidade e Formación Profesional da Xunta de Galicia. O noso agradecemento ós profesores da Universidade de Vigo, Emilio Fernández, codirector do proxecto e a Gloria Fiestras do Departamento de Estatística e Investigación Operativa polos seus sempre útiles consellos.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

ÍNDICES DE SOBOL NA ANÁLISE DE INFLUÊNCIA DAS PROPRIEDADES EM ESTRUTURAS COMPÓSITAS

Luísa N. Hoffbauer¹, Carlos C. António²

¹LAETA (INEGI), Instituto Superior de Engenharia do Porto, Politécnico do Porto

²LAETA (INEGI), Faculdade de Engenharia do Porto, Universidade do Porto

RESUMO

Vai ser estudado o problema da propagação de incertezas em estruturas compósitas laminadas. Primeiro é proposta uma abordagem baseada no projecto óptimo de estruturas com o objectivo de atingir um nível de fiabilidade específico e a correspondente carga máxima aplicada à estrutura compósita obtida em função do ângulo das camadas. Este cenário corresponde a um problema de fiabilidade inverso e é resolvido para os valores médios de referência das propriedades mecânicas dos compósitos laminados. Seguidamente, com o objectivo de hierarquizar as variáveis de entrada, gera-se um conjunto de pontos uniformemente dispersos num domínio centrado nos valores médios de referência das variáveis aleatórias de entrada, utilizando-se para o efeito o método do delineamento uniforme (*Uniform Design Method*, UDM). São então obtidos o número de Tsai, R , associado ao ponto de falha mais provável (*Most Probable Failure Point*, MPP), o índice de fiabilidade e as respetivas sensibilidades relativas para cada ponto obtido considerado no UDM. Em segundo lugar, usando os pontos gerados pelo UDM como padrões de entrada/saída, foi desenvolvida uma Rede Neuronal Artificial (RNA) com aprendizagem baseada em pesquisa evolucionária supervisionada. Em terceiro lugar, usando a RNA desenvolvida, é estudada a propagação de incertezas em função do ângulo das camadas utilizando os índices de Sobol.

Palavras e frases chave: Estruturas compósitas, Incertezas, Delineamento Uniforme, Fiabilidade, Rede Neuronal Artificial, Índices de Sobol .

1. INTRODUÇÃO

O uso de materiais compósitos laminados em todos os tipos de estruturas tem aumentado nos últimos anos. As variações nos parâmetros de fabrico, no controlo de qualidade e a heterogeneidade dos materiais, são alguns dos factores que introduzem incerteza nos materiais compósitos afectando o desempenho das aplicações estruturais destes materiais. Daí a necessidade de uma análise da propagação de incertezas que permita analisar a influência da propagação das incertezas na resposta estrutural e determinar como a incerteza na resposta estrutural pode ser atribuída às diversas variáveis de entrada, o que permite identificar as mais importantes.

2. PROPAGAÇÃO DE INCERTEZAS

O problema da propagação de incertezas de estruturas compósitas laminadas é estudado de acordo com os seguintes passos:

1. É considerada uma abordagem baseada no projecto óptimo de estruturas compósitas com o objectivo de atingir um índice de fiabilidade específico, β_a , sendo calculada a carga máxima correspondente em função do ângulo das camadas, a . Este problema de fiabilidade inverso é resolvido para os valores médios, μ_i , das propriedades mecânicas dos compósitos laminados.

2. Usando o método do delineamento uniforme (UDM) é gerado un conxunto de puntos pertencente ao intervalo $[\mu_i - \alpha\mu_i, \mu_i + \alpha\mu_i]$ cobrindo un dominio centrado nos valores médios de referencia das variábeis aleatorias. Este método permite unha exploración uniforme dos valores do dominio necesaria ao desenvolvemento do modelo da RNA.
3. Para cada punto do UDM son obtidos o número de Tsai crítico $\bar{R}$ asociado ao punto de falla máis provábel (MPP), o índice de fiabilidade β e as súas sensibilidades relativas tomando como referencia a carga máxima previamente calculada para o valor medio μ_i . O método de Lind-Hasofer é usado para a determinación do índice de fiabilidade. A análise de sensibilidades é realizada usando o método da variábel adjunta.
4. É desenvolvida una RNA usando una aprendizagem supervisionada baseada en pesquisa evolucionaria. Os puntos gerados polo UDM e os correspondentes valores resposta son usados como conxuntos de valores experimentais de entrada/saída no proceso de aprendizagem da RNA.
5. Usando a rede “ótima” desenvolvida é feita una análise global de sensibilidades usando os índices de Sobol.

Para un nivel de fiabilidade especificado, β_a , o problema de fiabilidade inverso é descrito pola ecuación

$$\begin{aligned} \min_{\lambda, a} [\beta_S(\lambda, a, \mathbf{x}) - \beta_a]^2 \\ \text{sujeito a} \\ 0 \leq a \leq \frac{\pi}{2} \end{aligned} \quad (1)$$

onde β_S é o índice de fiabilidade estrutural.

O vector das cargas aplicadas é definido por $\mathbf{L} = \lambda \mathbf{L}^{\text{ref}}$, onde $\mathbf{L}^{\text{ref}}$ é o vector carga de referencia. Para determinar o índice de fiabilidade estrutural β_S é aplicado o método de Lind-Hasofer e un esquema iterativo apropiado baseado en gradientes.

O número de Tsai, R_k , definido como sendo a razón entre a tensión máxima admissível e a tensión aplicada, é obtido do criterio de falla interactivo de Tsai-Wu calculado no k -ésimo punto da estrutura resolvendo a ecuación

$$1 - (F_{ij}s_i s_j) R_k^2 - (F_i s_i) R_k = 0 \quad i, j = 1, 2, 6 \quad (2)$$

onde s_i son as componentes do vector tensión e F_{ij} e F_i son os parámetros de resistencia asociados com os laminados reforçados unidireccionais definidos de un punto de vista macro-mecánico. A superficie estado limite $G(\mathbf{x})$ que separa o dominio de falla do dominio de seguración pode ser escrita na forma $G(\mathbf{x}) = \bar{R} - 1$ onde $\bar{R}$ é o número de Tsai crítico definido na ecuación

$$\bar{R} = \min(R_1, \dots, R_k, \dots, R_{N_s}) \quad (3)$$

sendo N_s o número total de puntos onde o vector tensión é calculado.

A localización destes puntos depende da metodoloxía do pós-processamento usada na análise estrutural. Nesta abordagem é usado o MEF e entón estes puntos coinciden com os puntos de Gauss da integración numérica.

2. DELINEAMENTO UNIFORME

Nos últimos anos têm sido propostos vários métodos de planeamento de experiencias para resolver problemas industriais. O obxectivo do planeamento de experiencias (*design of experiments*) é o de descubrir una relación entre a resposta e as variáveis de entrada, com o menor custo possível. A chave para este problema é determinar un conxunto de puntos (no dominio das variáveis de entrada) que permitam estimar de modo eficiente o valor esperado da resposta. Há dois tipos de abordagem para esta questão:

- Abordagem estocástica: do ponto de vista da estatística, pretende encontrar-se um planeamento tal que a média da amostra se um estimador não enviesado ou assintoticamente não enviesado da média real com a menor variância possível. Um exemplo desta abordagem é o método do Hipercubo Latino.
- Abordagem determinística: de um ponto de vista determinístico, pretende encontrar-se uma amostra tal que a diferença entre a média da amostra e a média real seja tão pequena quanto possível. A desigualdade de Koksma-Hlawka majora o erro do valor esperado da resposta. Esse majorante é o produto entre o valor da variação total da resposta e a discrepância do conjunto dos pontos no domínio das variáveis de entrada. Esta desigualdade indica que, quanto maior for a uniformidade dos pontos, menor é o erro. Interessa, pois, determinar pontos uniformemente dispersos no domínio das variáveis de entrada. Esta abordagem é designada por delineamento uniforme (*Uniform Design Method*, UDM).

Wang e Fang foram os primeiros a aplicar os métodos quasi-Monte Carlo ao planeamento de experiências, propondo uma então nova técnica, o delineamento uniforme. Esta técnica surgiu da necessidade de encontrar uma maneira de conseguir a máxima informação possível usando relativamente poucos pontos experimentais. A ideia chave do UDM é a de encontrar um conjunto de pontos que estejam uniformemente dispersos no hipercubo unitário s -dimensional C^s . Com efeito, o UDM pode ser considerado como uma versão do planeamento de experiências com o objectivo de minimizar a discrepância. Neste contexto, a discrepância é usada como uma medida de uniformidade de dispersão. Seja P um conjunto de n pontos em C^s . Para todo $\mathbf{g} = (g_1, g_2, \dots, g_s)$ seja $N(\mathbf{g}, P)$ o número de pontos contidos no hipercubo $[0, g_1] \times [0, g_2] \times \dots \times [0, g_s]$.

Então

$$D(n, P) = \sup_{\mathbf{g} \in C^s} \left| \frac{N(\mathbf{g}, P)}{n} - g_1 g_2 \dots g_s \right| \quad (4)$$

é chamada discrepância de P .

O quociente $\frac{N(\mathbf{g}, P)}{n}$ representa a percentagem de pontos que cai no hipercubo $[0, g_1] \times [0, g_2] \times \dots \times [0, g_s]$ enquanto $g_1 g_2 \dots g_s$ é o volume deste hipercubo. Como o volume total é unitário, $g_1 g_2 \dots g_s$ representa a percentagem de volume ocupada pelos n pontos. A medida $D(n, P)$ é um bom índice, pois é de esperar que, quanto maior for a uniformidade da dispersão no conjunto P , menor será a diferença entre $\frac{N(\mathbf{g}, P)}{n}$ e $g_1 g_2 \dots g_s$. Conforme a norma usada assim são obtidas várias medidas de discrepância.

O método do delineamento uniforme impõe que todos os pontos estejam uniformemente distribuídos em todas direcções, enquanto que no método do Hipercubo Latino os pontos estão uniformemente distribuídos em cada direcção.

A construção de um delineamento uniforme com n amostras e s factores é um problema complicado do ponto de vista computacional. Por isso, foram propostos métodos de construção como, por exemplo, o método *good lattice point* (glp). Baseado neste método, o UDM disponibiliza muitas tabelas que nos dão os valores no domínio das variáveis de entrada que deveremos considerar, de modo a obter pontos uniformemente dispersos nesse domínio. Uma tabela UDM é representada por $U_n(q^s)$, em que U denota o UDM, n o número de amostras, q o número de níveis de cada variável de entrada e s o número máximo de colunas da tabela. De notar que nestas tabelas podem ser escolhidas diferentes colunas, o que resulta em discrepâncias diferentes. Como resultado disto, para cada tabela do método do delineamento uniforme está associada uma tabela acessória que inclui recomendações relativamente às colunas que devem ser escolhidas para obter menor discrepância de acordo com o número de variáveis de entrada.

3. RNA DESENVOLVIDA

A RNA proposta é constituída por três camadas de nós (neurónios): a camada de entrada, a camada intermédia ou escondida e a camada de saída. A ligação entre os neurónios de entrada e os da camada intermédia e entre os da camada intermédia e os neurónios de saída são designadas por sinapses. Estas são ligações ponderadas que estabelecem a relação entre os dados de entrada, D_i^{inp} , e os dados de saída, D_j^{out} . Na RNA desenvolvida o vector dos dados de entrada, $\mathbf{D}^{inp}$, é definido por um conjunto de valores de variáveis aleatórias que são as propriedades mecânicas dos compósitos laminados tais como as propriedades elásticas ou de resistência. O módulo de

elasticidade longitudinal, E_1 , o módulo de elasticidade transversal, E_2 , a resistencia transversal em tracção Y e a resistencia ao corte, S são consideradas como variáveis de entrada da rede e são denotadas por $\mathbf{X} = [E_1, E_2, Y, S]$.

Nesta abordagem, cada conjunto de valores para o vector aleatório $\mathbf{X}$ é seleccionado usando o delineamento uniforme. O correspondente vector de dados de saída D^{out} contém o número de Tsai crítico, $\bar{R}$, o índice de fiabilidade estrutural, β_s e as sensibilidades relativas.

4. ÍNDICES DE SOBOL

Considerando que $\mathbf{X} = (X_1, \dots, X_n)$ são n variáveis aleatórias não correlacionadas e Ψ_m o funcional da resposta previamente definido, um indicador de importância da variável de entrada X_i pode ser baseado na variância de Ψ_m se se fixar X_i no seu verdadeiro valor x_i^* , isto é: $V(\Psi_m|X_i = x_i^*)$. Esta é a variância de Ψ_m condicionada a $X_i = x_i^*$. A variável X_i que leva a uma maior redução na variância de Ψ_m quando se fixa X_i no seu verdadeiro valor x_i^* , é a mais importante. No entanto, como na generalidade dos casos não é conhecido o verdadeiro valor x_i^* de X_i , ultrapassa-se esta dificuldade considerando a média da variância condicionada sob todas os possíveis valores x_i^* de X_i , ou seja, $E(V(\Psi_m|X_i))$, cujo valor é tanto maior quanto menos importante for X_i . Considerando a seguinte propriedade algébrica

$$V(\Psi_m) = V(E(\Psi_m|X_i)) + E(V(\Psi_m|X_i)) \quad (5)$$

verifica-se que ao menor valor de $E(V(\Psi_m|X_i))$ corresponde o maior valor de $V(E(\Psi_m|X_i))$.

O valor de $V(E(\Psi_m|X_i))$ pode ser usado como um indicador da importância de X_i na variância de Ψ_m . Este indicador é directamente proporcional à importância de X_i . O índice normalizado associado é

$$S_i = \frac{V(E(\Psi_m|X_i))}{V(\Psi_m)} \quad (6)$$

denominado índice de sensibilidade de primeira ordem de Sobol

5. EXEMPLO NUMÉRICO

Consideremos um painel compósito com a forma de uma asa de avião representado na Figura 1.

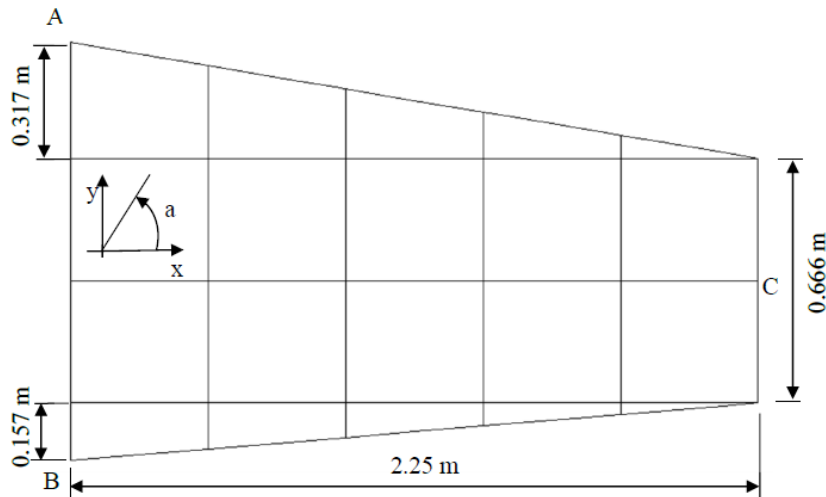


Figura 1: Definição geométrica do painel compósito com a forma de uma asa de avião.

A espessura do painel é de 0.015 m. A estrutura é encastrada no lado linear (AB) e livre nos restantes lados. É aplicada uma carga vertical com direcção perpendicular relativamente

ao plano OXY no ponto C. A estrutura é constituída por um laminado do sistema compósito carbono/resina de epóxico. Um laminado do tipo *angle-ply* balanceado, com oito camadas e a sequência de empilhamento $[+a/-a/+45^\circ/-45^\circ]_S$ numa construção simétrica. O ângulo da camada a é medido relativamente ao eixo dos xx do sistema de eixos de referência. Todas as camadas têm a mesma espessura. O elemento finito de casca é usado na análise estrutural. O índice de fiabilidade especificado é $\beta_a = 3$ e o coeficiente de variação de cada variável aleatória é fixado em $C.V.(X) = 6\%$, relativamente ao seu valor médio.

As variáveis aleatórias não correlacionadas são as propriedades mecânicas $X = (E_1, E_2, Y, S)$ com a distribuição seguinte $E_1 \sim \mathcal{N}(181.000, 10.860)GPa$, $E_2 \sim \mathcal{N}(10.300, 0.618)GPa$, $Y \sim \mathcal{N}(40.000, 2.400)MPa$, $S \sim \mathcal{N}(6.800, 4.080)MPa$.

A carga máxima obtida pela resolução do problema inverso formulado em 1 é calculada em função do ângulo das camadas e são obtidos os seguintes valores

Ângulo	0°	15°	30°	45°	60°	75°	90°
Carga máxima [N]	10412.5	9818.2	4785.6	2683.6	1298.15	1004.9	822.45

Tabela 1: Cargas de referência na análise da propagação de incertezas

Considere-se a Tabela $U_{27} (27^{10})$ do UDM onde são seleccionadas as colunas 1, 4, 6, 9 de acordo com o que é sugerido pela tabela acessória respectiva para quatro variáveis aleatórias e com discrepância $D(n, P) = 0.1189$. Os pontos definidos pelo UDM são considerados como valores experimentais de entrada a serem usados no processo de aprendizagem da RNA. Após a obtenção da RNA ótima para o painel compósito, as variáveis de entrada são ordenadas em função da sua importância baseada nos índices de Sobol (Tabelas 2 e 3).

Ordenação	0°	15°	30°	45°	60°	75°	90°
1°	Y	Y	Y	Y	Y	Y	Y
2°	E_2	E_2	S	E_1	E_2	E_1	E_2
3°	S	S	E_2	S	E_1	S	S
4°	E_1	E_1	E_1	E_2	S	E_2	E_1

Tabela 2: Ordenação dos parâmetros de entrada segundo a sua importância na resposta, $\bar{R}$, utilizando os índices de Sobol.

Ordenação	0°	15°	30°	45°	60°	75°	90°
1°	Y	Y	S	Y	Y	Y	Y
2°	E_2	S	Y	E_1	E_2	E_1	E_2
3°	E_1	E_2	E_2	S	E_1	E_2	E_1
4°	S	E_1	E_1	E_2	S	S	S

Tabela 3: Ordenação dos parâmetros de entrada segundo a sua importância na resposta, β , utilizando os índices de Sobol.

A variável aleatória mais importante na explicação da variabilidade do n° de Tsai é a resistência transversal, Y , para todo o domínio do ângulo das camadas. Também a resistência ao corte, S , é importante no intervalo $[15^\circ, 45^\circ]$. O módulo elástico longitudinal, E_1 , tem uma importância relevante no intervalo $[45^\circ, 75^\circ]$ e o módulo elástico transversal, E_2 , é importante em todo o domínio do ângulo das camadas a , excepto para 75° .

3. CONCLUSÕES

Foi realizado um estudo da influência da anisotropia na propagação de incertezas nos compósitos, baseado nas metodologias propostas.

Com base nos resultados numéricos, são determinadas e discutidas medidas de importância dos parâmetros de entrada na resposta estrutural, em função da anisotropia dos materiais compósitos.

A análise de propagação de incertezas é muito útil no projecto de estruturas compósitas laminadas, minimizando os efeitos indesejáveis das incertezas dos parâmetros de entrada na fiabilidade estrutural.

REFERÊNCIAS

- António, C. and Hoffbauer, L. (2010) Uncertainty propagation in inverse reliability-based design of composite structures. *International Journal of Mechanics and Materials in Design*, 6, 89–102.
- António, C. and Hoffbauer, L. (2013) Uncertain assessment approach for composite structures based on global sensitivity indices. *Composite Structures*, 99, 202–212.
- Fang, L. and Wang, Y. (1993) *Global Sensitivity Analysis*. Wiley.
- Saltelli, A. et al. (2008) *Global Sensitivity Analysis - The Primer*. John Wiley and Sons, Ltd.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

COVARIATE MEASUREMENT ERROR IN A NONLINEAR MODEL FOR LONGITUDINAL DATA

Santos Oliveira, Rosa

Matemática, Faculdade de Ciências; CINTESIS, Faculdade de Medicina; Instituto Politécnico de Bragança, Pólo de Vila Nova de Famalicão, Portugal

ABSTRACT

An usual aim in longitudinal studies is to characterise the relationship between a dichotomous outcome (death e.g) and both time-dependent and time-independent covariates. In this paper, we study nonlinear generalised mixed-effects regression models for analysis of longitudinal data. We indicate a two-stage approach that, in the first stage models a linear model to the longitudinal data to estimate the random effect that characterises the trajectory for each individual and in the second stage uses the estimated intercept and slope as predictors for the covariates used in a probit model. We show how to adapt a regression calibration and the best linear unbiased obtained from a linear mixed model approaches to this context and compare these approaches with a naive analysis where the estimation error is ignored. We show that the straight use of the estimates in the probit model produces biased estimates of their outcome. The methods are applied to health data collected during a study designed to evaluate the epidemiology of community-acquired sepsis in a larger cohort of infected intensive care unit patients. We show that the straight use of the estimates in the probit model produces biased estimates of their outcome. Nonetheless regression calibration and linear mixed effect offers little or no advantage when sample sizes are small, they performs best when samples are reasonably large and especially when the error of prediction (*measurement error*) or the effects are not small. Our study indicates that correcting for measurement errors instead of falsely assume that errors are not present will produce less bias than ignoring exposure measurement error in the analysis. Our study indicates that regression calibration and linear mixed effect method provides a way to obtain unbiased estimators, specially when is correction advocated, practically indistinguishable.

Keywords: error of prediction, general nonlinear model, longitudinal data

1. INTRODUCTION

Evaluating the diagnostic or prognostic ability of a biomarker usually focuses on a single measurement of a physiologic parameter and its association with an outcome of interest. However, the repeated measurement of the biomarker over a time period may offer additional information expressed as a trend or longitudinal profile of that parameter. For example Póvoa (2011) showed that in septic patients, the longitudinal profile of the biomarker C-reactive protein (CRP), summarized as the initial value (intercept) and rate of change per day (slope), was associated with the risk of death and could be used as a marker of sepsis resolution in the clinical setting.

The longitudinal trend of the biomarker may be summarized by a small number of parameters. In this article, we focus on a situation where the trend is linear and can be described by the intercept and slope. These parameters can then be used to study the association with the outcome of interest. In the sepsis example, this association was investigated by modelling the risk of death as a function of the intercept and slope of the linear profile of the CRP, using a logistic model.

Oliveira & Teixeira-Pinto (2019), have shown before how to use the ordinary least squares estimates of the intercept and slope to analyse their association with a dichotomous outcome, e.g. death, by framing the problem in the context of a nonlinear model with covariates measured with error. In particular, we showed that if the intercept and slope estimates were used as covariates without error in the prognostic/diagnostic logistic model, this could lead to significant bias but we could adapt the regression calibration (Carroll et al., 1996) and the joint modelling of the longitudinal modelling using a pseudo-likelihood (Fuller et al., 1987) - methods proposed in the

measurement error theory - as fitting strategies to reduce the bias. In this study we extend the previous approach by using the *best linear unbiased predictors (BLUPs)*, obtained from a linear mixed model as estimates of the linear profile of the biomarker, as covariates in a prediction model for a dichotomous outcome.

2. NOTATION AND MODEL SETUP

For the problem, the following three methods of estimating the linear measurement model coefficients are discussed: OLS, LME and RC.

We will concentrate on the problem of estimating the association between a true exposure and disease in a probit regression model, when exposure measurement is subject to bias and collected in a longitudinal study.

This section starts with the presentation of the covariate model containing distributional information of the latent variable and the error model relating the latent variable to the observable measurement (X) is described. Then the response model relating the true, but latent covariate (CRP trend and CRP baseline value) to the dependent variable death/survival state (D) in a longitudinal setting are presented. This section finish with the discussion on methods to estimate the parameters of the linear mixed effects model.

Let Y_{ij} be the j^{th} observed longitudinal measurement of the time-varying biomarker at time t_{ij} . Let D_i be a dichotomous response variable for the i^{th} subject in a sample of n individuals. We want to model D_i as a function of the longitudinal profile of Y_{ij} . We will contrast three approaches using a two step methodology, where we first we model the longitudinal profile on each individual and then, using estimates of this profile, we model the outcome D_i .

We assume a logistic regression structure:

$$prob(D_i = 1|X) = H(\alpha + \beta X), \quad (1)$$

where

$$H(x) = \frac{\exp(x)}{1 + \exp(x)}. \quad (2)$$

2.1 The longitudinal profile using Ordinary Least Squares

To model the longitudinal profile of each individual in a first step, linear regressions are fitted to each subject separately using ordinary least squares (OLS), where Y_{ij} is the dependent variable and the time t_{ij} is the only covariate.

$$Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \epsilon_{ij}, \quad (3)$$

where the $\epsilon_{ij} = N(0, \sigma_\epsilon^2)$ is the random effects residual deviations assumed to be distributed independently, β_{0i} represents the initial value of the biomarker and β_{1i} is the rate of change over time, for the i^{th} individual.

In the second step, we use the OLS estimates of the initial value and rate of change of the biomarkers for each individual, $\hat{\beta}_{OLS,0i}$ and $\hat{\beta}_{OLS,1i}$, for the model of the binary response D_i . We choose a probit regression for convenience of calculations, although the logit link would also be a common option.

$$pr(D_i = 1|\hat{\beta}_{OLS,i}) = \Phi(\alpha_0 + \hat{\beta}_{OLS,i}^T \alpha + \mathbf{Z}_i^T \gamma) = \Phi(\alpha_0 + \alpha_1 \hat{\beta}_{OLS,0i} + \alpha_2 \hat{\beta}_{OLS,1i} + \mathbf{Z}_i^T \gamma). \quad (4)$$

$\mathbf{Z}_i$ represents a vector of observations for a set of additional covariates that we might want to include in the model for the outcome D_i , and γ the respective regression coefficients. The parameters α_0 , α_1 , α_2 and γ are estimated using the maximum likelihood estimators.

2.2 The longitudinal profile using a Linear Mixed Model

Rather than modelling the biomarker's longitudinal trajectory of each individual separately as before, we can model all the profiles simultaneously by using a linear mixed model with a random intercept, representing the subject specific initial value of the biomarker, and a random effect for time, representing the individual rate of change per unit of time, i.e.,

$$Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim \text{Normal}(0, \sigma_\epsilon^2) \quad (5)$$

where β_{0i} represents the initial value of the biomarker and β_{1i} its rate of change over time. Additionally, we suppose that, unconditionally,

$$\beta_i = (\beta_{0i}, \beta_{1i})^T = \text{Normal}\{\beta = (\beta_0, \beta_1)^T, \Sigma_\beta\}.$$

We can now compute the BLUP for β_{0i} and β_{1i} -- $\hat{\beta}_{BLUP,0i}$ and $\hat{\beta}_{BLUP,1i}$ -- and use these quantities as estimates of the individual initial value and slope. These estimates are then treated as fixed covariates on the probit model for the binary response D_{ij} .

$$\text{pr}(D_i = 1 | \hat{\beta}_i) = \Phi(\alpha_0 + \hat{\beta}_{i,BLUP}^T \alpha + \mathbf{X}_i^T \gamma) = \Phi(\alpha_0 + \alpha_1 \hat{\beta}_{BLUP,0i} + \alpha_2 \hat{\beta}_{BLUP,1i} + \mathbf{X}_i^T \gamma). \quad (6)$$

As before, $\mathbf{X}_i$ represents additional covariates of interest in the probit model. The parameters α_0 , α_1 , α_2 and γ can now be estimated using the maximum likelihood estimators. As in the previous case, when using the OLS estimates for the intercepts (β_{0i}) and slopes (β_{1i}), this approach is an approximation given that the BLUPs are treated as fixed covariates and the uncertainty of their estimation, i.e., their prediction is ignored in the model for the binary response.

2.3 Regression Calibration

By recognising that using estimates for the intercept and slope in models 3 and 4 is similar to fitting a model with covariates measured with error, we can apply established and computationally efficient methodologies available in measurement error theory. One of this methods is the regression-calibration (RC) approach. The basic idea of RC is to regress the outcome not on the covariate with error but on the expected value of the error free covariate given the covariate error. In our case, instead of fitting the model for D_i using the estimates of β_0 and β_1 , we fit the model using the $E(\beta | \hat{\beta})$.

Considering Σ_{ui} the covariance matrix of the error associated with the estimation of β , we have derived before that if the model for

$$\text{pr}(D_i | \hat{\beta}_i, z) = \Phi \left[\frac{\alpha_0 + E(\beta_i | \hat{\beta}_i)^T \alpha + \mathbf{Z}^T \gamma}{\{1 + \alpha^T \text{cov}(\beta_i | \hat{\beta}_i) \alpha\}^{1/2}} \right]. \quad (7)$$

where

$$E(\beta_i | \hat{\beta}_i) = \beta + \Sigma_\beta (\Sigma_\beta + \Sigma_{ui})^{-1} (\hat{\beta}_i - \beta)$$

and

$$\text{cov}(\beta_i | \hat{\beta}_i) = \Sigma_\beta \{I - (\Sigma_\beta + \Sigma_{ui})^{-1} \Sigma_\beta\}.$$

the logistic regression coefficients can then be adjusted for bias due to measurement error in a one-step procedure to obtain estimates of α and γ .

Finally, since $\int \Phi(a + bx) \phi(x) dx = \Phi\{a/(1 + b^2)^{1/2}\}$, it is easily seen that on an approximation of the likelihood function where the basic idea is to plug in "best" predictions for the covariates

measured with error and go ahead with the estimation of GLMs by adjusting point and interval estimates of effect obtained from regression models as if these predictions were covariates measured without error. In regression calibration the adjusted value is the expectation of the true value conditional on the measured value. Suppose X is measured with error and W is observed. In regression calibration we substitute the conditional expectation of X on W , $X_{RC}(W) = E(X|W)$, and run the regression model $E(Y|W, \beta) \approx f\{X_{RC}(W), \beta\}$.

For the new method we will first assume measurement error in X such that $E(W|Y) = E(X|Y)$. In other words, W is considered to be an unbiased measurement of X .

Carroll et al. (1996) version of RC is used, when first we fit the MEM to the data to produce a estimate of the coefficients β_i , via OLS. Then we fit the logistic regression model:

$$Pr(D_i = 1|Y_i) = H(\beta_0 + \beta_1 Y_i), i = 1, \dots, n \quad (8)$$

where $H(\cdot)$ is defined in 2 by replacing Y_i by its estimated conditional expectation.

3. APPLICATION TO SACIUCI DATA

Examination of the effect of CRP (a biomarker) on the prognosis of death for patients with SEPSIS. The idea is to model CRP trajectory as a function of time allowing to model longitudinal trajectories and the variability associated with subject-specific differences and measurement. The intercept describes the "initial" CRP value and the slope describes the CPR rate of change per day for a specific patient. Estimation is conducted to predict ability of C-Reactive Protein (CRP) in Sepsis resolution, analysing longitudinal changes in SACiUCI data. The three proposed procedures are applied to the SACiUCI a prospective, cohort, multi-centred study, conducted over one year (1 December, 2004 to 30 November, 2005) of biomarkers as prognostic factors in SEPSIS (Póvoa et al., 2002). A major interest of this study is to estimate the relationship between serial CRP measurements after prescription of antibiotics and the clinical course of Community-Acquired Sepsis (SAC) admitted in Intensive Care Units (ICU). Data of community-acquired sepsis study was collected in Seventeen intensive care units (ICU) in Portugal. All adult patients (age > 18 years) consecutively admitted in the participating ICUs were enrolled and screened for CAS. Sepsis and sepsis-related conditions were defined according to the criteria proposed by the American College of Chest Physicians/Society of Critical Care Medicine (bone, 1992). A total of 891 patients were enrolled from 1 December 2004 to 30 November 2005. Only patients with more than 2 CRP measurements were taken into account in the analysis, a total of 713 patients were considered. Patients were evaluated at baseline and at each of the five following days respecting the variables of interest. All admitted patients were followed up until death or hospital discharge. Data on various factors and variables were recorded for each patient. The focus of this analysis is on the rate of change of CRP during the five days, as well CRP baseline value. Age also was investigated as variable of interest to predict the discharge status of survival/death. The clinical suggestion is that the decrease of CRP over time may be an indicator of sepsis resolution.

Assuming a linear growth model with subject-specific random intercepts an slope, we have $y_{ij} = \beta_{0i} + t_{ij}\beta_{1i} + \beta_{2i} + \Sigma_i$.

In this model (age-adjusted), β_0 and β_1 (fixed) treatment-specific intercept and slope and β_2 age used for adjustment. Different intercepts and slopes for the CRP are assumed. The intercept and slopes indicates CRP baseline value (measured at time $t = 0$) and the change of the CRP over time, respectively. Therefore, the estimation and comparison of the CRP slopes is of primary interest. The parameter β_2 subject-specific random intercept. Inclusion of β_2 allows the CRP change for a given subject to have its own intercept. To allow the change curve to have both subject-specific intercept β_{20} and slope β_{21} the linear model is

The dichotomous outcome of the SACiUCI study only distinguishes between the states of death ($Y = 1$) and survival ($Y = 0$). The results are shown in Table 1. The regression coefficient estimates using the regression calibration and the GLS methods (Naive 2) in subjects for whom all 5 CRP measurements where made are identical.

4. CONCLUSION

Táboa 1: Regression Estimates from SACiUCI data.

Estimator	Estimate	SE	OR	LL	UL
<i>OR</i> _{95%CI}					
CPR init (α_1)					
<i>Naive</i>	0.099	0.253	1.576	(0.960,	2.588)
<i>Naive2</i>	0.105	0.270	1.554	(0.916,	2.636)
<i>RegCal</i>	0.105	0.270	1.554	(0.916,	2.636)
CPR init (α_2)					
<i>Naive</i>	0.292	0.747	13.115	(3.031,	56.755)
<i>Naive2</i>	0.330	0.845	18.187	(3.474,	95.197)
<i>RegCal</i>	0.330	0.845	18.187	(3.474,	95.197)
CPR init (<i>age</i>)					
<i>Naive</i>	0.004	0.011	1.041	(1.019,	1.062)
<i>Naive2</i>	0.004	0.011	1.041	(1.019,	1.062)
<i>RegCal</i>	0.004	0.011	1.041	(1.019,	1.062)

It is well known that *measurement errors* in the predictors bias the estimates, and for that reason several methods have been proposed to solve this problem. However, our work presents a new direction in this field, as it is new in logistic/probit regression models with *measurement error* and binary response, where the *measurement error* distribution does not need to be known neither replicate data need to be available. We have compared three methods performance in both realistic and extreme settings using simulation.

The three methods discussed in this paper assume nondifferential *measurement error*, which is required for valid application of regression calibration method. Nevertheless, this assumption is usually satisfied in prospective studies.

Our simulations show that regression calibration produces parameter estimates that are "practically" indistinguishable from those produced by Linear Mixed Effect model. Parameters estimation for both RC and LME had excellent performance.

Compared to Ordinary least squares, RC and LME offers little or no advantage when sample sizes are small, but performs best when samples are reasonably large and especially when the *measurement error* or the effects are not small.

From the analysis presented in this paper, it is possible to derivate some curious results.

First, note that the naive approach produces weak results.

The regression calibration model produces the same results as the linear mixed effect model. In fact, both the LME and the RC perform well and give good estimates, typically requiring less than 10 Fisher-Scoring iterations to maximize the likelihood. However, in the case of one variable, it is more efficient to use the OLS (linear model) and with more variables linear mixed effect model, but always with the correction advocated.

There remain a number of unanswered questions about correction for covariate measurement error in nonlinear mixed effects models when applied to longitudinal data or repeated measures. For instance, when the number of covariates is greater then the number of measurements and for so, deserve further study. One of our, nowadays, research aims.

REFERENCES

- Bone R.C.; Balk R.A.; Cerra F.B.; Dellinger R.P.; Fein A.M.; Knaus W.A.; Schein R.M. and Sibbald W.J. (1992) Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee, American College of Chest Physicians/Society of Critical Care Medicine, 101, 1644-1655.
- Carroll, R. J.; Küchenhoff, Helmut; Lombard, F and Stefanski, L. A. (1996) Asymptotics for the SIMEX estimator in nonlinear measurement error models. Journal of the American Statistical Association, 91, 242-250.
- Fuller, W. (1987) Measurement Error Models. John Wiley and Sons, 1987.
- Póvoa, P. (2002) C-reactive protein: a valuable marker of sepsis. Intensive Care Medicine, 28 (3), 235-243.

Póvoa, P.; Teixeira-Pinto A. M.; Carneiro; A. H. and Portuguese Community-Acquired Sepsis Study Group SACiUCI (2011). C-reactive protein, an early marker of community-acquired sepsis resolution: a multi-center prospective observational study. *Critical Care*, 15 (4).

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

UNHA ANÁLISE NON PARAMÉTRICA SOBRE A EFICIENCIA E A PRODUTIVIDADE DOS PAÍSES EUROPEOS EN TECNOLOXÍA E PROPIEDAD INTELECTUAL

Agostinho Rosa¹; Armando B Mendes²; Jorge Santos³ e Luis.Cavique⁴

¹ Universidade de Évora, Escola de Ciências Sociais, Departamento de Economia, Centro de Estudos e Formação Avançada em Gestão e Economia – CEFAGE-UE, arosa@uevora.pt.

² Universidade dos Açores, NIDeS e ALGORITMI armando.b.mendes@uac.pt

³ Universidade de Évora, Esc. de Ciências e Tecnologia, Dep. de Matemática, Instituto de Investigação e Formação Avançada Centro de Investigação em Matemática e Aplicações, jmas@uevora.pt

⁴ Universidade Aberta, BioISI-MAS luis.cavique@uab.pt

RESUMO

O obxectivo dos autores foi aplicar a análise de envoltura de datos para estudar a eficiencia da ciencia e da investigación en 31 países europeos. A análise de envoltura de datos é unha técnica moi ben establecida no estudo da eficiencia de sistemas complexos con varios inputs e outputs. A ese respecto recolléronse datos da base de datos de Eurostat. Segundo os datos dispoñibles, o último ano completo e consolidado foi 2012, polo que escollemos este ano para o estudo.

Os principais resultados son que baixo o presuposto de constante retorno de escala só dous países eran eficientes: Suíza e Suecia. Cando se permiten retornos variables á escala Alemaña e dous países pequenos: Malta e Islandia tamén se consideran eficientes.

Palabras e frases chave: Análise de envoltura de datos; Análise de eficiencia; Programación lineal; Ciencia, tecnoloxía e innovación; Produtividade.

1. INTRODUCCIÓN

A tecnoloxía e a innovación científicas son variables moi importantes no desenvolvemento do progreso hoxe en día. Os autores utilizaron os datos de 2012 dispoñibles de Eurostat para avaliar a eficiencia de 31 países europeos.

O modelo que empregaron baseouse no traballo seminal de Farrel en 1957 (Farrel; 1957) bem como no modelo CCR desenvolvido por Charnes Cooper e Rhodes (Charnes et al, 1978) e Banker Charnes e Cooper en 1984 (Banker et al, 1984). Este modelo de Banker Charnes e Cooper aplícase baixo a suposición de retornos variables de escala.

Os autores buscaron os datos máis recentes na base de datos, pero o ano máis recente dispoñible foi 2012.

Consideraron como consumos (Inputs):

1 Poboación total (en miles).

2 HRST=Persoas con educación terciaria e/ou empregadas en ciencia e tecnoloxía (en miles).

Consideraron como resultados (Outputs):

1 As solicitudes de patentes europeas de alta tecnoloxía dirixidas á Oficina Europea de Patentes,

2 As solicitudes de patentes europeas de baixa tecnoloxía dirixidas á Oficina Europea de Patentes,

Sempre utilizamos o modelo de maximización de resultados xa que é máis razoable maximizar as solicitudes de patentes europeas en vez de reducir os insumos, xa que son claramente non discretos e non podemos reducilos.

Primeiro asumimos que a asunción de retornos constantes á escala era válida e obtivemos tan eficientes só dous países. Despois permitimos que os retornos variables a escala xa que o tamaño dos países son moi diferentes con países grandes como Alemaña e pequenos como Malta e Islandia.

2. ANÁLISE DE DATOS E RESULTADOS

Primeiro houbo unha análise conxunta dos resultados a nivel europeo, que se presenta en figura 1 con valores que van desde o 0,08 ao máis de 165,13.

Dada a especificidade deste congreso, a figura 2 mostra un mapa con só os valores de Portugal e España que oscilan entre os 22,9 e os 42,8.

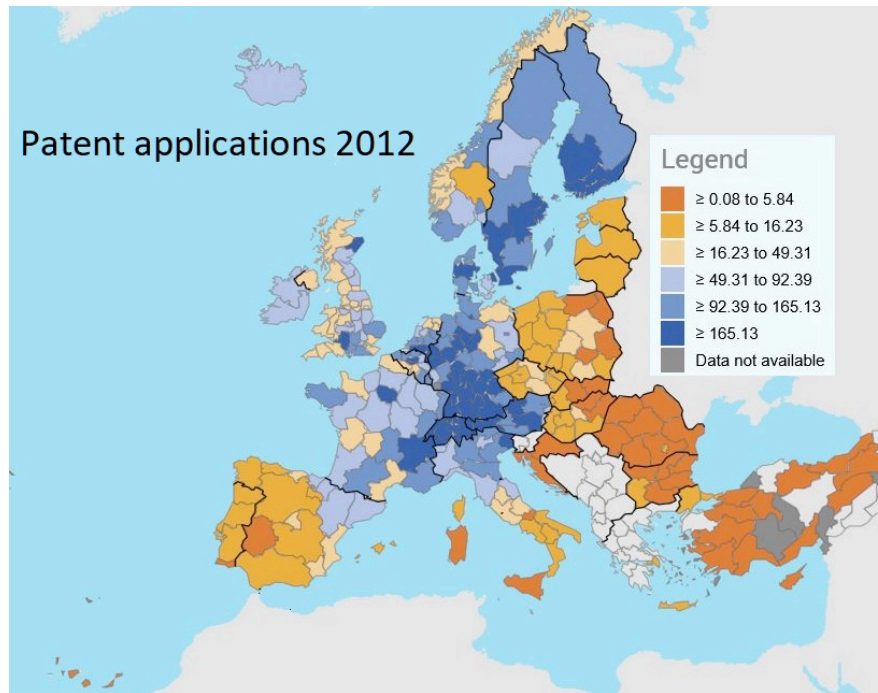


Figura 1: Solicitudes de patentes á Oficina Europea de Patentes (OEP) no ano 2012 [TSC00009]
Fonte de datos: Eurostat

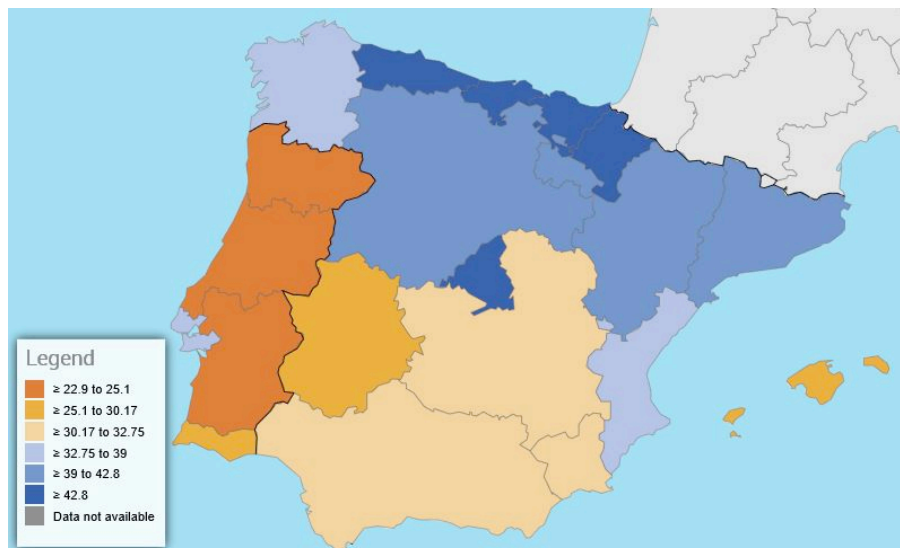


Figura 2: Solicitudes de patentes á Oficina Europea de Patentes (OEP) no ano 2012 ES+PT
Fonte de datos: Eurostat

A figura 3 é unha captura de pantalla que ten a folia de cálculo que será lida polo software EMS coa súa convención habitual {I} é unha entrada e {O} significa que a variable é unha saída (Scheel;2000).

Liechtenstein foi excluído xa que faltaban valores nalgúns variables.

	Pop {I}	Per {I}	High {O}	Low {O}
BE	11076	2875	361	1146
BG	7327	1307	6	28
CZ	10505	2163	21	211
DK	5581	1548	231	1087
DE	80328	21223	2863	19067
EE	1325	382	5	18
IE	4589	1257	85	229
EL	11086	1933	16	86
ES	46818	10623	293	1225
FR	65277	15770	2070	6825
HR	4276	665	1	19
IT	59394	9671	384	3949
CY	862	237	1	2
LV	2045	455	8	20
LT	3004	711	16	16
LU	525	166	7	60
HU	9932	1835	53	155
MT	418	75	0	6
NL	16730	4726	709	2680
AT	8408	1886	232	1631
PL	38064	7242	60	424
PT	10542	1620	15	97
RO	20096	2503	16	56
SI	2055	474	10	116
SK	5404	1018	8	37
FI	5401	1569	462	1176
SE	9483	2794	891	2194
UK	63495	18165	1185	4204
IS	320	86	1	26
NO	4986	1502	83	440
CH	7955	2569	370	2872

Figura 3: Datos básicos para o noso estudo

Na figura 4 presentamos os resultados dos dous modelos que consideramos: CCR e BCC a partir das siglas dos seus autores respectivamente Charnes Cooper & Rhodes e Banker Charnes & Cooper.

Notamos como era de esperar que as ineficiencias dos modelos VRS (BCC) nunca superen as dos modelos CRS (CCR).

C:\00u\p\lems\Vigo.xls_CRS_RAD_OUT											C:\00u\p\lems\Vigo.xls_VRS_RAD_OUT										
DMU	Score	Pop (I)	Per (I)	High (O)	Low (O)	Benchmarks	(S) Pop (I)	(S) Per (I)	(S) High (O)	(S) Low (O)	DMU	Score	Pop (I)	Per (I)	High (O)	Low (O)	Benchmarks	(S) Pop (I)	(S) Per (I)	(S) High (O)	(S) Low (O)
1 BE	218.50%	0.000	2.185	0.374	0.626	27 (0.76) 31 (0.29)	1537.01	0.00	0.00	0.00	1 BE	216.49%	0.000	1.779	0.358	0.642	5 (0.01) 27 (0.75) 31 (0.24)	39.62	0.00	0.00	0.00
2 BG	609.25%	0.000	46.093	0.301	0.699	27 (0.20) 31 (0.29)	3111.24	0.00	0.00	0.00	2 BG	488.52%	0.000	47.256	0.303	0.697	27 (0.20) 29 (0.53) 31 (0.27)	85.83	0.00	0.00	0.00
3 CZ	143.97%	0.000	11.440	0.000	1.000	31 (0.84)	3806.99	0.00	76.27	0.00	3 CZ	138.60%	0.000	11.726	0.000	1.000	29 (0.16) 31 (0.84)	98.19	0.00	75.57	0.00
4 DK	141.13%	0.000	1.411	0.287	0.713	27 (0.21) 31 (0.37)	612.67	0.00	0.00	0.00	4 DK	138.50%	0.000	1.447	0.289	0.711	27 (0.21) 29 (0.43) 31 (0.36)	91.43	0.00	0.00	0.00
5 DE	120.49%	0.000	1.205	0.222	0.778	27 (0.61) 31 (7.38)	3339.82	0.00	0.00	0.00	5 DE	25.80%	0.000	0.000	0.000	1.000		7			
6 EE	898.32%	0.000	18.589	0.360	0.640	27 (0.09) 31 (0.05)	65.94	0.00	0.00	0.00	6 EE	577.57%	0.000	19.048	0.362	0.638	27 (0.08) 29 (0.89) 31 (0.03)	27.35	0.00	0.00	0.00
7 IE	449.32%	0.000	4.493	0.413	0.587	27 (0.41) 31 (0.05)	357.52	0.00	0.00	0.00	7 IE	436.21%	0.000	4.602	0.415	0.585	27 (0.40) 29 (0.56) 31 (0.04)	35.40	0.00	0.00	0.00
8 EL	320.42%	0.000	23.204	0.261	0.739	27 (0.19) 31 (0.55)	4945.18	0.00	0.00	0.00	8 EL	299.11%	0.000	23.798	0.263	0.737	27 (0.19) 29 (0.27) 31 (0.54)	31.13	0.00	0.00	0.00
9 ES	829.90%	0.000	8.299	0.312	0.688	27 (1.84) 31 (2.14)	2395.55	0.00	0.00	0.00	9 ES	590.98%	0.000	3.885	1.000	0.000	5 (0.42) 27 (0.58)	38.13	0.00	0.00	21.30
10 FR	203.97%	0.000	2.040	0.365	0.635	27 (3.99) 31 (1.80)	3126.19	0.00	0.00	0.00	10 FR	110.12%	0.000	0.915	1.000	0.000	5 (0.70) 27 (0.30)	11.93	0.00	0.00	58.45
11 HR	936.98%	0.000	39.370	0.000	1.000	31 (0.26)	2217.49	0.00	75.94	0.00	11 HR	654.65%	0.000	40.355	0.000	1.000	29 (0.77) 31 (0.23)	76.16	0.00	68.91	0.00
12 IT	273.79%	0.000	2.738	0.000	1.000	31 (3.76)	9450.12	0.00	338.63	0.00	12 IT	229.86%	0.000	2.126	0.000	1.000	5 (0.38) 31 (0.62)	86.23	0.00	39.05	0.00
13 CY	545.70%	0.000	75.457	1.000	0.000	27 (0.08)	59.27	0.00	0.00	72.57	13 CY	143.03%	15.087	63.726	1.000	0.000	18 (0.24) 27 (0.06) 29 (0.70)	0.00	0.00	0.00	67.08
14 LV	866.28%	0.000	18.663	0.424	0.576	27 (0.16) 31 (0.01)	506.14	0.00	0.00	0.00	14 LV	629.23%	0.000	19.366	0.659	0.341	18 (0.32) 27 (0.14) 29 (0.54)	31.66	0.00	0.00	0.00
15 LT	406.92%	0.000	14.069	1.000	0.000	27 (0.25)	590.67	0.00	0.00	327.08	15 LT	293.98%	0.000	14.455	1.000	0.000	18 (0.77) 27 (0.23)	64.53	0.00	0.00	05.04
16 LU	306.21%	0.000	3.062	0.000	1.000	31 (0.06)	12.40	0.00	2.44	0.00	16 LU	168.04%	3.332	0.000	0.243	0.757	27 (0.00) 29 (0.97) 31 (0.03)	0.00	13.24	0.00	0.00
17 HU	999.06%	0.000	9.991	0.395	0.605	27 (0.55) 31 (0.12)	3795.52	0.00	0.00	0.00	17 HU	966.88%	0.000	10.234	0.397	0.603	27 (0.54) 29 (0.34) 31 (0.12)	81.52	0.00	0.00	0.00
18 MT	514.47%	0.000	15.145	0.000	1.000	31 (0.03)	186.96	0.00	10.72	0.00	18 MT	big	0.011	532.384	0.000	1.000		3			
19 NL	163.29%	0.000	1.633	0.334	0.666	27 (0.98) 31 (0.78)	1282.48	0.00	0.00	0.00	19 NL	149.14%	1.193	0.000	0.365	0.635	5 (0.10) 27 (0.82) 31 (0.07)	0.00	34.61	0.00	0.00
20 AT	126.66%	0.000	1.267	0.213	0.787	27 (0.05) 31 (0.68)	2530.31	0.00	0.00	0.00	20 AT	125.44%	0.000	1.300	0.214	0.786	27 (0.05) 29 (0.28) 31 (0.67)	14.83	0.00	0.00	0.00
21 PL	876.25%	0.000	18.762	0.211	0.789	27 (0.16) 31 (2.65)	5508.49	0.00	0.00	0.00	21 PL	635.63%	0.000	14.841	0.000	1.000	5 (0.25) 31 (0.75)	80.42	0.00	17.98	0.00
22 PT	790.31%	0.000	17.903	0.231	0.769	27 (0.08) 31 (0.54)	5456.02	0.00	0.00	0.00	22 PT	762.34%	0.000	18.366	0.232	0.768	27 (0.08) 29 (0.39) 31 (0.53)	35.00	0.00	0.00	0.00
23 RO	052.79%	0.000	40.528	0.352	0.648	27 (0.58) 31 (0.34)	1861.76	0.00	0.00	0.00	23 RO	044.44%	0.000	41.532	0.354	0.646	27 (0.58) 29 (0.08) 31 (0.34)	58.26	0.00	0.00	0.00
24 SI	454.54%	0.000	4.545	0.000	1.000	31 (0.18)	589.35	0.00	21.52	0.00	24 SI	404.18%	0.000	4.659	0.000	1.000	29 (0.84) 31 (0.16)	43.86	0.00	17.39	0.00
25 SK	757.27%	0.000	27.573	0.281	0.719	27 (0.13) 31 (0.28)	2146.33	0.00	0.00	0.00	25 SK	645.27%	0.000	28.274	0.282	0.718	27 (0.13) 29 (0.54) 31 (0.24)	14.58	0.00	0.00	0.00
26 FI	106.34%	0.000	1.063	0.427	0.573	27 (0.54) 31 (0.02)	90.80	0.00	0.00	0.00	26 FI	104.36%	0.000	1.089	0.429	0.571	27 (0.53) 29 (0.45) 31 (0.02)	73.64	0.00	0.00	0.00
27 SE	90.98%	0.910	0.000	1.000	0.000	22					27 SE	66.50%	0.341	0.000	1.000	0.000		21			
28 UK	391.56%	0.000	3.916	0.349	0.651	27 (4.14) 31 (2.57)	3803.73	0.00	0.00	0.00	28 UK	202.08%	1.491	0.000	1.000	0.000	5 (0.76) 27 (0.24)	0.00	20.87	0.00	62.98
29 IS	363.10%	0.000	3.631	0.000	1.000	31 (0.03)	53.90	0.00	7.51	0.00	29 IS	big	568.761	0.410	0.996	0.004		18			
30 NO	349.65%	0.000	3.496	0.264	0.736	27 (0.15) 31 (0.42)	208.75	0.00	0.00	0.00	30 NO	342.94%	0.000	3.586	0.265	0.735	27 (0.15) 29 (0.44) 31 (0.41)	86.02	0.00	0.00	0.00
31 CH	65.74%	0.657	0.000	0.000	1.000	27					31 CH	64.18%	0.659	0.000	0.000	1.000		20			

Input Output Data C:\00u\p\lems\Vigo.xls

Figura 4: Resultados dos dous modelos

3. CONCLUSIONES

Só se atoparon dous países eficientes no primeiro modelo, Suíza a costa do seu bo rendemento en patentes de baixa tecnoloxía e Suecia debido ao seu bo rendemento en patentes de alta tecnoloxía.

No segundo modelo que considera rendementos variables a escala, Alemaña, Malta e Islandia fixéronse eficientes ao ser penalizados no primeiro modelo polo seu tamaño.

Malta e Islandia conducen a programas lineais sen solución admisible xa que teñen os valores máis baixos en Poboación total e Persoas con educación terciaria e/ou empregadas en ciencia e tecnoloxía respectivamente.

Tamén se poderían aplicar técnicas de restrición de peso para evitar que algunhas unidades teñan pesos nulos ou no outro extremo moi altos.

4. AGRADECEMENTOS

Os autores agradecen as súas respectivas universidades e centros de investigación pola súa colaboración.

REFERENCIAS

Banker RD, A Charnes and WW Cooper (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science* 30(9): 1078-1092.

Charnes A, WW Cooper and EL Rhodes (1978). "Measuring the Efficiency of Decision Making Units." *EJOR* 2: 429-444.

FARRELL, M. J., (1957) "The Measurement of Productive Efficiency," *J. Roy. Statist. Soc. Ser. A*, 120 , 253-290

Scheel, H. (2000) EMS: Efficiency Measurement System User's Manual.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

NUEVOS MODELOS DERIVADOS DEL TRUCK AND TRAILER ROUTING PROBLEM

Laura Davila Pena¹ y Balbina V. Casas Méndez¹

¹IMAT, Instituto de Matemáticas, MODESTYA, Grupo de Modelos de Optimización, Decisión, Estadística y Aplicaciones, Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela.

RESUMEN

Los problemas de rutas de vehículos, VRPs (Dantzig y Ramser, 1959), admiten diversas variantes, según las necesidades de los clientes y otras características que se pueden presentar al analizar situaciones de la vida real. Una de ellas es el denominado truck and trailer routing problem, TTRP (Chao, 2002), donde una flota de camiones y tráileres atienden a un conjunto de clientes tal que cuando un tráiler no puede acceder a un cliente, este es atendido solo por un camión.

El objetivo de este trabajo es presentar un nuevo modelo, derivado del TTRP, que incorpora vehículos divididos en compartimentos: el MC-TTRP. Este problema está motivado por Guitián de Frutos y Casas-Méndez (2019), que considera la distribución de piensos en una cooperativa agraria en la cual los socios demandan productos que no se pueden mezclar. Se implementa este modelo en AMPL y, usando datos de un problema real, se resuelve con Gurobi.

Palabras e frases chave: problemas de transporte, modelos de rutas de vehículos, compartimentos, optimización lineal entera mixta.

Agradecimientos: La investigación de Laura Davila Pena está financiada por el Ministerio de Educación, Cultura y Deporte del Gobierno de España (contrato FPU17/02126). Las autoras agradecen la financiación recibida del FEDER (MINECO/AEI proyecto MTM2017-87197-C3-3-P) y de la Xunta de Galicia (Grupo Competitivo ED431C 2017/38).

REFERENCIAS

- Chao, I. M. (2002) A tabu search method for the truck and trailer routing problem. *Computers & Operations Research*, 29, 33–51.
- Dantzig G. B. and Ramser, J. H. (1959) The truck dispatching problem. *Management Science*, 6, 80–91.
- Guitián de Frutos, R. M. and Casas-Méndez, B. V. (2019) Routing problems in agricultural cooperatives: A model for optimization of transport vehicle logistics. *IMA Journal of Management Mathematics*. To appear.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

ESTIMACIÓN DEL CORE-CENTER EN JUEGOS DE AEROPUERTO

Nadine Espinoza Burgos¹ y Estela Sánchez Rodríguez²

¹ Grupo SiDOR. Universidade de Vigo.

² Departamento de Estatística e Investigación Operativa. Universidade de Vigo.

RESUMEN

El problema del aeropuerto, introducido por Littlechild y Owen (1973), es un problema clásico de asignación de costes. ¿Cómo se reparten los costes entre las distintas compañías de un modo justo? A cada problema se le puede asociar un juego cooperativo con utilidad transferible, de esta forma, las soluciones conjunto y los repartos puntuales del juego asociado proporcionan repartos de los costes entre las aerolíneas.

Existen distintas soluciones que atienden a diferentes principios. Entre ellas está el core-center, González-Díaz y Sánchez-Rodríguez (2007), que promedia de forma ecuánime todos los repartos estables del núcleo. En González, Mirás, Quinteiro y Sánchez (2016) se obtienen expresiones explícitas del core-center para el juego del aeropuerto. Una de las expresiones desarrolladas es el cociente de volúmenes, específicamente el core-center se corresponde con la razón entre el volumen del core del juego con un jugador clonado y el volumen del core del juego original.

El cálculo de volúmenes exactos es un problema NP-Hard, por lo que en este trabajo estudiamos posibles estimaciones de esta solución para dimensiones elevadas. Partimos de la esencia del core-center, tratando a todos los jugadores de la misma forma, para utilizar varios métodos de aproximación, desde los tradicionales como el método de aceptación y rechazo o el método de inversión, hasta el método de caminos aleatorios hit-and-run (Smith (1984), Emiris y Fisikopoulos (2018)). En todas las estimaciones se utiliza el método de Monte Carlo con el objetivo de mejorar la aproximación. Describiremos las ventajas e inconvenientes de los métodos antes descritos y estudiaremos el error de las estimaciones.

Palabras e frases clave: Juego del aeropuerto, core-center, aproximación, volumen politopo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por the European Regional Development Fund (ERDF) y por el Ministerio de Economía, Industria y Competitividad a través del proyecto MTM2017-87197-C3-2-P y por la Xunta de Galicia a través del Fondo Europeo de Desarrollo Regional (Grupos de Referencia Competitiva ED431C-2016-040).

REFERENCIAS

- Cao Abad, R. (2002) Introducción a la simulación y a la teoría de colas. Netbiblo.
- Emiris, I.Z. y Fisikopoulos, V. (2018). Practical polytope volume approximation. *ACM Transactions on Mathematical Software*, 44, n. 4.
- González Díaz, J., Mirás Calvo, M. A., Quinteiro Sandomingo, C. y Sánchez Rodríguez, E. (2016) Airport games: The core and its center. *Mathematical Social Sciences*, 82, 105–115.
- González Díaz, J. y Sánchez Rodríguez, E. (2007) A natural selection from the core of a tu game: the core-center. *International Journal of Game Theory*, 36, 1, 27–46.
- Littlechild, S.C. y Owen, G (1973). A simple expression for the Shapley value in a special case. *Management Science*, 20, 370–372.
- Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32, 6, 1296–1308.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

ALGORITMOS PARA EL CÁLCULO DEL CORE-CENTER DE UN JUEGO DE BANCARROTA

Iago Núñez Lugilde¹ y Estela Sánchez Rodríguez²

¹Grupo SiDOR. Universidade de Vigo

²Universidade de Vigo

RESUMEN

Un problema de bancarrota, O'Neill (1982), surge cuando un recurso no es suficiente para satisfacer todas las demandas de un conjunto de acreedores. En Aumann y Maschler (1985) se presenta un análisis de este problema utilizando juegos cooperativos y se introducen diferentes soluciones a dicho problema. Thomson (2003, 2015) son buenas referencias recopilatorias de la ingente literatura que ha ido apareciendo a lo largo de los años para esta clase de problemas. En González-Díaz y Sánchez-Rodríguez (2007) se introduce el core-center como solución de un juego cooperativo con núcleo no vacío, y se define como el centro de gravedad del núcleo. Posteriormente, Mirás Calvo, Quinteiro Sandomingo y Sánchez-Rodríguez (2018) introducen un algoritmo para la obtención del core-center del juego de bancarrota y estudian sus principales propiedades. El algoritmo proporciona una teselación finita del conjunto de imputaciones en la que el núcleo del juego de bancarrota es una de las piezas de la descomposición.

En este trabajo desarrollamos otro algoritmo a partir de una teselación interior del núcleo del juego bancarrota, a partir de núcleos de juegos de bancarrota básicos, que son aquellos en los que el estado es igual a la semisuma de las demandas o bien todos los acreedores demandan más que el estado. En estos juegos, las soluciones más conocidas coinciden en el reparto propuesto. En el primer caso, cada agente recibe exactamente la mitad de su demanda y en el segundo el estado es dividido a partes iguales.

Compararemos ambos algoritmos para calcular el core-center. Veremos que el algoritmo exterior es más eficiente en tiempo computacional cuánto mayor sea la complejidad del núcleo del juego en relación al número de puntos extremos del mismo, mientras que el algoritmo interior será más eficiente cuánto menor sea la complejidad del núcleo.

Teniendo en cuenta el análisis anterior, presentamos un algoritmo combinación de ambos, que aprovecha las ventajas de los dos algoritmos y que nos permite calcular el core-center de un problema de bancarrota de forma exacta con un elevado número de jugadores y de forma aproximada cuando no sea posible la computación exacta.

Palabras y frases clave: Problema de bancarrota, juego de bancarrota, núcleo de un juego, core-center, teselación, algoritmo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por the European Regional Development Fund (ERDF) y por el Ministerio de Economía, Industria y Competitividad a través del proyecto MTM2017-87197-C3-2-P y por la Xunta de Galicia a través del Fondo Europeo de Desarrollo Regional (Grupos de Referencia Competitiva ED431C-2016-040).

REFERENCIAS

- Aumann, R.J. y Maschler, M. (1985). Game theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36, 195–213.
- González Díaz, J. y 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.

- Mirás Calvo, M. Á., Quinteiro Sandomingo, C. y Sánchez Rodríguez, E. (2018). The core-center of a bankruptcy problem. Mimeo.
- Thomson, W., (2003). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: a survey. *Mathematical Social Sciences* 45, 249—297.
- Thomson, W. (2015). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: An update. *Mathematical Social Sciences* 74, 41–59.
- O'Neill, B. (1982). A problem of rights arbitration from the Talmud. *Mathematical Social Sciences* 2, 345–371.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

OPTIMIZACIÓN DE RUTAS DE CARROS MEZCLADORES EN UNA COOPERATIVA AGRÍCOLA

Carlos Amiama¹, Luisa Carpenté² y Silvia Lorenzo-Freire²

¹Universidade de Santiago de Compostela, Departamento de Producción Vegetal y Proyectos de Ingeniería.

²Universidade da Coruña, grupo MODES, CITIC.

RESUMEN

En este trabajo se aborda el problema de optimizar automáticamente las rutas de trabajo de una serie de carros mezcladores de una cooperativa agrícola que deben visitar a diario una serie de explotaciones ganaderas. El objetivo principal es minimizar el tiempo de trabajo total de estas máquinas durante un horizonte temporal dado. Además, se han de cumplir una serie de restricciones relacionadas con la jornada laboral, hora de llegada a las explotaciones, equilibrio de trabajo entre las máquinas, etc. Para resolver el problema, se ha diseñado un modelo de programación lineal entera. El modelo se ha aplicado a unos datos reales, obteniendo finalmente una comparativa entre la solución propuesta y la hasta ahora vigente en la cooperativa agrícola.

Palabras y frases clave: Aplicaciones de la programación lineal entera, optimización de rutas.

1. INTRODUCCIÓN

Debido a la liberalización y globalización de los mercados, especialmente el agrícola, cobra gran importancia el aumento de la eficiencia en cualquiera de los procesos involucrados en este sector con el fin de alcanzar productos competitivos. Este trabajo se centra en el problema real de una cooperativa agrícola gallega. Concretamente, se trata de diseñar, de manera automática y eficiente, las rutas de unos carros mezcladores que dicha cooperativa posee. Los carros mezcladores son unas máquinas que se encargan de preparar la mezcla con la que se alimenta el ganado en las explotaciones.

En la literatura de optimización, se encuentran algunas referencias a problemas de rutas que surgen en cooperativas agrícolas. Por ejemplo, en Blanco et al. (2010) se diseñan rutas de forma automática para cosechadoras empleando un algoritmo heurístico. En ese trabajo, no era posible calcular la solución mediante el modelo de programación matemático propuesto en un tiempo razonable. Una situación similar se puede consultar en Amiama et al. (2015), donde se emplean técnicas heurísticas para el diseño de rutas eficientes de recogida de leche.

La situación concreta que analizamos en este trabajo, parte de tres carros mezcladores que están localizados en un almacén y que trabajan en una misma zona de actuación. Cada carro ha de visitar una serie de explotaciones diariamente e invertir un tiempo en realizar la mezcla correspondiente. El tiempo de mezcla en una misma explotación puede variar en días consecutivos. Esto se debe a que hay explotaciones que tienen vacas de lactación, que requieren la mezcla diariamente, y también novillas, que requieren la mezcla cada dos días.

Para solucionar este problema, se ha diseñado un problema de programación lineal entera multiperiodo (dos días) con los siguientes elementos:

Objetivo 1.

Se ha de buscar la combinación de rutas que minimice los tiempos de desplazamiento de los carros mezcladores.

Objetivo 2.

La carga de trabajo diaria ha de ser similar. Se persigue asimismo, que los tiempos de trabajo de cada carro mezclador sean parecidos.

Restricciones.

- Los carros parten diariamente del almacén y finalizan su jornada diaria en dicho almacén.
- Todas las explotaciones han de ser visitadas diariamente.
- Cada carro visita las mismas explotaciones cada día.
- El tiempo de llegada a cada explotación no debe variar mucho de un día para otro.
- Cada carro no puede sobrepasar su jornada laboral de nueve horas.

La solución obtenida con este modelo se ha comparado con la solución manual que venía aplicando el técnico de la cooperativa.

REFERENCIAS

Amiama, C., Pereira, J. M., Carpenté, L., & Salgado, J. (2015). Spatial decision support system for the route management for milk collection from dairy farms. *Transportation Letters*, 7(5), 279-288.

Blanco, V., Carpenté, L., Hinojosa, Y., & Puerto, J. (2010). Planning for agricultural forage harvesters and trucks: model, heuristics, and case study. *Networks and Spatial Economics*, 10(3), 321-343.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

MEJORANDO LA MOTIVACIÓN DE TUS ESTUDIANTES CON TÉCNICAS DE INNOVACIÓN EDUCATIVA

Marta Cousido-Rocha¹, María Gómez-Rúa²

¹Departamento de Estadística e IO, y Grupo de investigación SiDOR, Universidade de Vigo, Facultade de CC Económicas e Empresariais, Campus Lagoas-Marcosende, 36310 Vigo, España.

²Departamento de Estadística e IO, y Grupo de investigación ECOSOT, Universidade de Vigo, Facultade de CC Económicas e Empresariais, Campus Lagoas-Marcosende, 36310 Vigo, España.

RESUMEN

Como docentes, día a día buscamos ofrecer una enseñanza de calidad. Para ello es fundamental que nuestros alumnos/as se involucren en el aprendizaje de una manera activa. Las técnicas de innovación educativa pueden ayudar a lograrlo. En este trabajo hemos estudiado los beneficios de dos técnicas de innovación docente: Pedagogía Inversa y Gamificación. Los resultados obtenidos han sido altamente satisfactorios. Los alumnos/as respondieron a dichas técnicas con un incremento en su motivación y consecuentemente con una mayor participación en clase, lo que culminó en una mejora del aprendizaje.

Palabras y frases clave: Innovación Educativa; Motivación; Gamificación; Pedagogía Inversa.

1. INTRODUCCIÓN

En este trabajo describimos nuevas metodologías para la innovación docente, tanto dentro como fuera del aula. Dichas técnicas combinan la pedagogía y la tecnología. Por supuesto, sin renunciar a la resolución de problemas, a la creatividad del alumno, y al fomento de la cooperación. Durante el curso 2018/2019 aplicamos este tipo de metodologías de innovación educativa en la materia Estadística de la titulación Programación Conjunta de Enseñanzas Oficiales (en adelante, PCEO) Grado en Administración y Dirección de Empresas/Grado en Derecho.

2. TÉCNICAS DE INNOVACIÓN EDUCATIVA

En esta sección presentamos varias técnicas que fomentan la innovación en la docencia así como nuestra experiencia con ellas y las tareas concretas llevadas a cabo en la materia anteriormente especificada.

Una técnica que tiene un enorme potencial es la denominada **Pedagogía Inversa**. Esta metodología didáctica consiste en transferir ciertas tareas que habitualmente se realizan en clase al trabajo en casa. De esta forma, en el aula se realizan actividades que requieren más participación e interacción, y en casa, a través de soportes audiovisuales, se trabajan algunos ejercicios que antes se realizaban en clase, ver O'Flaherty y Phillips (2015).

Otra técnica de aprendizaje es la **Gamificación**, que consiste en la utilización del juego para motivar el aprendizaje, creando así situaciones de estudio divertidas y proporcionando una mejora de la participación de los estudiantes (ver Lee y Hammer, 2011). A través de esta técnica aprovechamos la predisposición natural de los jóvenes hacia el mecanismo de los juegos para mejorar la motivación, la adquisición de conocimientos y el desarrollo de competencias.

En el curso 2018/19 hemos empleado estas dos metodologías en la materia Estadística de la titulación PCEO Grado en Administración y Dirección de Empresas/Grado en Derecho. Siguiendo la Pedagogía Inversa, se les proporcionó a los alumnos/as material previamente a la sesión magistral (aproximadamente dos días antes) y se les dieron indicaciones para preparar la clase con el mismo. El material consistió en vídeos introductorios, claros y cortos sobre conceptos que

posteriormente se explicarían en clase. Además de estos vídeos, los alumnos/as dispusieron de una serie de preguntas o ejercicios breves para poder poner a prueba la comprensión del vídeo. En la Figura 1 podemos ver como ejemplo la actividad propuesta para trabajar la sección “Contrastes de hipótesis” del tema “Conceptos y principios metodológicos de la inferencia estadística”.

Inferencia Estadística: Contrastes de Hipótesis

En este documento se describe la actividad propuesta para iniciar la comprensión de conceptos básicos presentes en un contraste de hipótesis:

1. Primeramente os proponemos ver los siguientes vídeos sobre contrastes de hipótesis que hemos seleccionado. Todos ellos son muy breves y explican los conceptos de una forma clara y intuitiva:
 - Introducción a los contrastes de hipótesis: <https://www.youtube.com/watch?v=nTDu8UUsjzQ&feature=youtu.be>.
 - Tipos de error en un contraste de hipótesis: https://www.youtube.com/watch?v=hnhONQ_H1OU&feature=youtu.be.
 - Definición de p -valor: <https://www.youtube.com/watch?v=RD0GqLJ0nkE&feature=youtu.be>.
2. Tras ver con detenimiento los vídeos anteriores debemos dar respuesta a las siguientes preguntas (todas la respuestas se encuentran en los vídeos):
 - ¿Qué es una hipótesis estadística?
 - ¿Qué tipos de hipótesis estadísticas tenemos?
 - ¿Cuáles son los resultados posibles de un contraste de hipótesis?
 - La hipótesis de un contraste, ¿se refiere a la población o a la muestra?
 - ¿Qué es la hipótesis nula en un contraste?
 - ¿Qué es la hipótesis alternativa en un contraste?
 - ¿Qué es el error tipo I?
 - ¿Qué es el error tipo II?
 - ¿Qué valores del p -valor llevan a rechazar la hipótesis nula?
 - ¿Cuándo se dice que un contraste es significativo?

Figura 1: Pedagogía Inversa. Ejemplo de una de las actividades propuestas a lo largo del curso la cual aborda la sección “Contrastes de hipótesis” del tema “Conceptos y principios metodológicos de la inferencia estadística”.

En el ámbito de la Gamificación hemos empleado el servicio web Kahoot! Mediante esta aplicación hemos creado diferentes cuestionarios con preguntas de la materia. La puesta en práctica fue la siguiente: los alumnos/as instalaron en su dispositivo móvil (u ordenador) la aplicación, a continuación el profesor utilizó un proyector para que pudieran seguir el cuestionario diseñado en forma de juego. Los alumnos/as respondieron a cada pregunta a través de su dispositivo, y al final de cada una de ellas podían ver la respuesta correcta, cuántos alumnos/as habían fallado y cuáles eran los tres compañeros con mejor puntuación. La aplicación premia tanto el hecho de que la respuesta sea correcta como la rapidez del estudiante en responder. También empleamos la Gamificación para ayudar a los alumnos/as con el aprendizaje de los conceptos de la sección “Contrastes de hipótesis” del tema “Conceptos y principios metodológicos de la inferencia estadística”. En la Figura 2 se pueden ver algunas de las preguntas de dicho cuestionario junto a la visualización del mismo una vez se inicia el juego.

Qué es una hipótesis estadística

Contrastando una hipótesis

53

0 Answers

Una suposición sobre la población

Una suposición sobre la muestra

Una suposición sobre la técnica estadística a utilizar

Una suposición sobre la variable estadística a utilizar

kahoot.it Game PIN: 185903

Figura 2: Gamificación, Kahoot. Ejemplo de uno de los Kahoots propuestos a lo largo del curso la cual aborda la sección “Contrastes de hipótesis” del tema “Conceptos y principios metodológicos de la inferencia estadística”.

3. CONCLUSIONES

Hemos comprobado cómo gracias a la Pedagogía Inversa, los alumnos/as llegaban a clase preparados y con conocimientos previos muy útiles para aprovechar al máximo todas las explicaciones del docente en el aula. Por otra parte la Gamificación nos permitió hacer una valoración de los conocimientos adquiridos durante la sesión de una manera lúdica. A través del Kahoot! los alumnos/as veían el cuestionario como un momento de descanso y juego mientras que en realidad seguían aprendiendo. La valoración general de la puesta en práctica de ambas técnicas fue muy positiva, tanto por parte de las profesoras como por parte de los estudiantes. Se observó dinamismo en clase y un aumento de la motivación por parte de los estudiantes en cuanto a la materia se refiere.

REFERENCIAS

- O’Flaherty, J. y Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, 25, 85-95.
- Lee, J. y Hammer, J. (2011). Gamification in Education: What, How, Why Bother?. *Academic Exchange Quarterly*, 15, 1-5.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

**PLUGIN DE R-COMMANDER PARA LA ENSEÑANZA DE ESTADÍSTICA
BÁSICA: RcmdrPlugin.TeachStat**

T. R. Cotos Yáñez¹, M. A. Mosquera Rodríguez¹, A. Pérez González¹, B. Reguengo Lareo

¹Departamento de Estatística e Investigación Operativa, Universidade de Vigo

RESUMEN

En este trabajo mostramos la motivación que nos llevó a la creación de un plugin de R-Commander que se adapte más específicamente a las necesidades docentes de la materia de Estadística en varias de las titulaciones de grado existentes en el Campus de Ourense, Universidade de Vigo. También se muestran las principales características del paquete señalando las ventajas más destacables tanto desde el punto de vista del docente como del alumno.

Palabras clave: Estadística, R, R-Commander, docencia.

*XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019*

¿AVANZAN LOS PAÍSES HACIA EL LOGRO DE LOS OBJETIVOS ESTABLECIDOS EN POLÍTICAS AMBIENTALES? ANÁLISIS MULTIVARIANTE DE UNA DÉCADA.

Fernández-Gómez, M.J.^{1,2}, Martín-Vallejo, J.^{1,2}, Barrera-Mellado, I.^{1,2},
Gallego-Alvarez, I.³, y Sastre-Hernández, B.¹

¹ Departamento de Estadística. Universidad de Salamanca.

² IBSAL. Universidad de Salamanca.

³ Departamento de Administración y Economía de la Empresa. Universidad de Salamanca.

RESUMEN

En este trabajo se estudia, mediante técnicas de Análisis Multivariante de tablas Multivía, la evolución del Índice de Desempeño Ambiental (EPI) desde 2008 hasta 2018. Se analiza la estabilidad de las relaciones entre los distintos indicadores que construyen el EPI y cómo estos pueden explicar las similitudes y diferencias entre los 61 países del mundo para los que se dispone de información completa a lo largo de la serie. Mediante la técnica X-STATIS se ha identificado la estructura común de covariación entre los indicadores de las áreas a lo largo de esta década. Se han encontrado dos periodos de tiempo claramente diferenciados en cuanto a las relaciones entre indicadores, entre países, y entre indicadores-países. Se observa cómo con el paso de los años gran parte de los países no sólo mantienen su interés en aumentar la salud ambiental sino que también ha crecido su preocupación sobre la protección y calidad de los ecosistemas para supervisar su vitalidad y llegar a un equilibrio sostenible.

Palabras y frases clave: EPI (Índice de desempeño ambiental), Política ambiental, Sostenibilidad, X-STATIS

REFERENCIAS

- Emerson, J.W., Esty, D.C., Levy, M.A., Kim, C.H., Mara, V., de Sherbinin, A. And Srebotnjak, A. (2010) Environmental Performance Index. New Haven, C: Yale Center for Environmental Law and Policy.
- Esty, D.C. and Porter, M. (2005) National environmental performance: an empirical analysis of policy results and determinants. *Environmental and Development Economics*, 10, 391-434.
- Esty, D.C., Levy, M.A., Kim, C.H., de Sherbinin, A., Srebotnjak, T., and Mara, V. (2008) Environmental Performance Index. New Haven: Yale Center for Environmental Law and Policy.
- Jafrenou, P.A. (1978) Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. Thèse 3^o cycle. Université des Sciences et Technique du Languedoc, Montpellier-II.
- Thioulouse, J. and Chessel, D. (1987) Les analyses multitableaux en Écologie Factorielle. 1. De la théorie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica. Oecologia generalis*, 8, 463-480.
- Blanc, L., Chessel, D. and Dolèdec, S. 1998. Étude de la stabilité temporelle des structures spatiales par analyses d'une série de tableaux de relevés faunistiques totalement appariés. *Bulletin Français de la Pêche et de la pisciculture*, 348, 1-21

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

A GENERALIZED PRODUCT-LIMIT ESTIMATOR OF THE SURVIVAL FUNCTION WHEN CURES ARE PARTIALLY KNOWN

Wende Clarence Safari¹, M. Amalia Jácome² and Ignacio López-de-Ullibarri³

¹Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, 1507 A Coruña, Spain. email: wende.safari@udc.es

²MODES, CITIC, Department of Mathematics, Faculty of Science, Universidade da Coruña, Spain. email: maria.amalia.jacome@udc.es.

³MODES, Department of Mathematics, Escuela Universitaria Politécnica, Ferrol, Universidade da Coruña, Spain. email: ignacio.lopezdeullibarri@udc.es.

ABSTRACT

We consider the survival function for univariate right censored event time data when a cure fraction is present. This means that there are subjects under study that will never experience the event of interest. Under random right censoring, the cure status of the censored times is unknown. In this context, the conditional survival function of the lifetime is estimated non-parametrically with the generalized product-limit estimator by Beran (1981). We consider the case when some censored individuals are observed to be cured given certain prior information e.g. the result of a diagnostic test or when their survival time exceeds a known threshold. We propose a nonparametric estimator for the conditional survival function in this situation, that reduces to the usual estimator in the general context without known cures. We prove that the estimator is consistent and asymptotically normally distributed. A bootstrap procedure is proposed to select the bandwidth. We show the good performance of the estimator with simulations. For illustration, the method is applied to a sarcoma dataset from the University Hospital of Santiago de Compostela, Spain.

Keywords: Bandwidth; Bootstrap; Censoring; Cure models; Kernel estimator; Nadaraya-Watson weights.

1. INTRODUCTION

Typical survival analyses assume that all individuals will eventually experience the event of interest. Cure models are used when this supposition is not reasonable and instead assume that the population of individuals is made up of two distinct groups: those who will experience the event of interest and those who will not. Some examples can be found in health science (e.g., time to the death from cancer, some patients will never die because of cancer), in economy (e.g., on duration of unemployment, part of the population may never get a job) or sociological studies (e.g., on the age at which a person gets married, a percentage of people will stay unmarried), among many other fields. Recent reviews of cure models can be found in Peng and Taylor (2014), and Amico and Van Keilegom (2018), among others.

Let Y be the time to the event, and X a covariate that might have an impact on the variable Y . We set $Y = \infty$ if the subject is cured. The probability of being cured when the covariate X equals x is $1 - p(x) = P(Y = \infty | X = x)$, and the conditional survival function of the uncured individuals given the covariate x , also called latency, is $S_0(t|x) = P(Y > t | Y < \infty, X = x)$. The mixture cure model, introduced by Boag (1949), assumes that the population is split into two groups, the cured and the susceptible to the event. As a consequence, the conditional survival function of Y , $S(t|x) = 1 - F(t^-|x) = P(Y > t | X = x)$, can be written as:

$$S(t|x) = 1 - p(x) + p(x)S_0(t|x).$$

The interest of the estimation of the conditional survival function with a cure fraction is evident, since both the cure rate and the latency can be estimated nonparametrically using the corresponding nonparametric estimator of $S(t|x)$, just considering the following relationships:

$$1 - p(x) = \lim_{t \rightarrow \infty} S(t|x) > 0 \text{ and } S_0(t|x) = \frac{S(t|x) - (1 - p(x))}{p(x)}.$$

When analyzing time-to-event data, it is very frequent that the lifetime of some subjects is censored, that is, the follow-up finishes before the event occurs. Under right random censoring, the conditional survival function $S(t|x)$ can be estimated fully nonparametrically with the generalized product-limit estimator proposed by Beran (1981). Under the assumption that the lifetime random variable Y is independent of the random censoring variable C given the covariate X , the asymptotic properties of the generalized product-limit estimator can be found in Dabrowska (1987, 1989, 1992), Li and Doss (1995), González-Manteiga and Cadarso-Suárez (1994) and Iglesias-Pérez and González-Manteiga (1999), among others. The generalized product-limit estimator in Beran (1981) has been recently used to derive completely nonparametric estimators of the cure rate (Xu and Peng, 2014; López-Cheda et al, 2017a) and the latency function (López-Cheda et al, 2017b).

With a cure fraction, only the uncensored observations are known to be uncured, since the event is observed to happen. But censoring prevents from observing whether a censored subject would experience the event or not, hindering the classification of the censored observations in cured or uncured. It is customary to assume no additional information on the cure status, thus, to model it as a latent variable. However, there might be situations when extra information about the cure status is available. For example, for some diseases an individual might be assumed to be cured if the observed survival time is greater than a cure threshold, or based on a diagnostic test.

Only few authors have explored cure models which allow cure to be partially known. Laska and Meisner (1992) proposed nonparametric estimators of the cure rate and latency when cure status is known if an individual does not develop an event of interest within a pre-specified time period. However, they did not incorporate covariates. Betensky and Schoenfeld (2001) proposed nonparametric estimation methods in the presence of random cure times and competing risk. They defined hospital discharge as the cure event and death as a competing event, but likewise did not incorporate covariates. Nieto-Baraja and Yin (2008) proposed a Bayesian semi-parametric approach for estimating a survival data with a cure fraction in the presence of covariates. Wu et al (2014) extended the Cox proportional hazards cure model in the possibility of knowing the cure status for an individual based on a diagnostic test with a given sensitivity and specificity. More recently, Bernhardt (2016) proposed a cure rate model which accommodates different censoring distributions for the cured and uncured groups, and allows for some individuals to be observed as cured when their survival time exceeds a known threshold. The model is developed for a general case where any parametric distribution can be used to model $S_0(t|x)$ and an expectation-maximization (EM) algorithm is provided for estimating the model parameters. He also showed by means of simulations that ignoring a cure threshold when one exists can lead to biased estimates.

None of the aforementioned works considered a completely nonparametric cure model in the presence of known cure fraction and covariates. This motivates us to introduce a nonparametric conditional survival function in the presence of known cure, that could be the departure point for further nonparametric estimation of the cure rate and the latency function.

Our estimator is an extension of the estimator in Beran (1981) to the case when cure is partially known. We describe it and give some properties in the next section. For the choice of the bandwidth, we recommend the use of a selector based on the bootstrap.

2. METHODOLOGY

Let Y be the time to event of interest, possibly depending on a covariate X , and C a right censoring random time with conditional distribution function $G(t|x)$. Furthermore, it is assumed that C is independent of Y given X . Let $T = \min(Y, C)$ be the observed time with distribution function $H(t|x)$, $\delta = 1(Y \leq C)$ the uncensoring indicator, and $\nu = 1(Y = \infty)$ the cure indicator.

Suppose that extra information about the cure status is provided. Then, let ξ be a binary variable indicating whether the cure status is known ($\xi = 1$) or not. We assume that cure status ν is independent of Y , C and X . So, the observed data $\{(X_i, T_i, \delta_i, \xi_i, \xi_i \nu_i), i = 1, \dots, n\}$ can be classified into

three groups: (a) the individual is observed to have experienced the event and therefore known to be uncured ($X_i, T_i = Y_i, \delta_i = 1, \xi_i = 1, \xi_i \nu_i = 0$); (b) the lifetime is censored and the cure status is missing ($X_i, T_i = C_i, \delta_i = 0, \xi_i = 0, \xi_i \nu_i = 0$); and (c) the lifetime is censored and the individual is known to be cured ($X_i, T_i = C_i, \delta_i = 0, \xi_i = 1, \xi_i \nu_i = 1$).

When cure is partially known, the conditional cumulative hazard function can be estimated nonparametrically as follows:

$$\hat{\Lambda}_h^c(t|x) = \sum_{i=1}^n \frac{B_{h[i]}(x) \delta_{[i]} \mathbf{1}(T_{(i)} \leq t)}{\delta_{[i]} B_{h[i]}(x) + \sum_{j=i+1}^n B_{h[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 0) + B_h^c(x)}$$

where $T_{(i)}$'s are the ordered observed times $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$, and $X_{[i]}$, $\delta_{[i]}$, $\xi_{[i]}$ and $\nu_{[i]}$ the corresponding concomitants,

$$B_{h[i]} = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^n K_h(x - X_{[j]})} \quad (1)$$

are the Nadaraya-Watson weights with $K_h(u) = \frac{1}{h} K(\frac{u}{h})$ the rescaled kernel and bandwidth $h > 0$, and

$$B_h^c(x) = \sum_{j=1}^n B_{h[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 1) \quad (2)$$

is the sum of the weights of the individuals known to be cured.

When cure is partially known, a nonparametric estimator of the conditional survival function $S(t|x)$ is

$$\hat{S}_h^c(t|x) = 1 - \hat{F}_h^c(t|x) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}(T_{(i)} \leq t)}{\delta_{[i]} B_{h[i]}(x) + \sum_{j=i+1}^n B_{h[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 0) + B_h^c(x)} \right), \quad (3)$$

with B_h and B_h^c given in (1) and (2), respectively.

The proposed nonparametric estimator of the survival function in (3) has got the following general properties.

1. If there are not known cures, $\hat{S}_h^c(t|x)$ reduces to the generalized product-limit estimator in Beran (1981):

$$\hat{S}_h(t|x) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}(T_{(i)} \leq t)}{\sum_{j=i}^n B_{h[j]}(x)} \right).$$

2. If there is no censoring, then $\hat{S}_h^c(t|x)$ reduces to the kernel estimator of the conditional survival function:

$$\tilde{S}_h(t|x) = \sum_{i=1}^n B_{h[i]}(x) \mathbf{1}(Y_i > t).$$

3. In an unconditional setting then $\hat{S}_h^c(t|x)$ reduces to

$$\hat{S}_n^c(t) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} \mathbf{1}(T_{(i)} \leq t)}{1 + \sum_{j=i+1}^n \mathbf{1}(\xi_{[j]} \nu_{[j]} = 0) + m} \right),$$

where m is the number of subjects known to be cured. In the particular case where an individual is known to be cured only if the observed time is greater than a known fixed time, say d , then the estimator reduces to the generalized maximum likelihood estimator in Laska and Meisner (1992):

$$\hat{S}_n^{LM}(t) = \prod_{i=1}^n \left(1 - \frac{\delta_{[i]} \mathbf{1}(T_{(i)} \leq t)}{\delta_{[i]} + (n_1 - i) + m} \right),$$

where $n = n_1 + m$ with n_1 the number of observed times smaller than d .

4. If there exists a common specific known cure threshold, say d , then $\hat{S}_h^c(t|x)$ reduces to the generalized product-limit estimator in Beran (1981).

3. ASYMPTOTIC PROPERTIES

We present some asymptotic properties of the estimator of the conditional survival function $S(t|x)$, showing its consistency and asymptotic normality. Before we state our results, we introduce some notation and a list of technical assumptions. Consider the functions

$$\begin{aligned} H(t|x) &= P(T \leq t | X = x) \\ H^1(t|x) &= P(T \leq t, \delta = 1 | X = x) \end{aligned} \quad (4)$$

$$\begin{aligned} H^{11}(t|x) &= P(T \leq t, \xi = 1, \nu = 1 | X = x) \\ J(t|x) &= 1 - H(t|x) + H^{11}(t|x) \end{aligned} \quad (5)$$

To prove the asymptotic properties, the following assumptions will be assumed throughout:

(A1) X, Y and C are absolutely continuous random variables.

(A2) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of the density function of X , $m(x)$, and $I_\delta = [x_1 - \delta, x_2 + \delta]$ for some $\delta > 0$ such that

$$0 < \gamma = \inf\{m(x) : x \in I_\delta\} < \sup\{m(x) : x \in I_\delta\} = \Gamma < \infty$$

and $0 < \delta\Gamma < 1$. For all $x \in I$ the random variable Y is independent of C given $X = x$.

(b) There exist $a, b \in \mathbb{R}$ satisfying

$$\inf\{J(t|x) : x \in I_\delta\} \geq \theta > 0 \text{ for } (t, x) \in [a, b] \times I_\delta.$$

(A3) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A4) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A5) The first derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times I_\delta$.

(A6) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times I_\delta$.

(A7) The derivatives first with respect to x and second with respect to t of $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded for $(t, x) \in [a, b] \times I_\delta$.

(A8) The densities (subdensity) corresponding to the distribution (subdistribution) functions $H(t|x)$ and $H^1(t|x)$ are bounded away from 0 in $[a, b]$.

(A9) The kernel function K is a symmetrical density vanishing outside $(-1, 1)$ and the total variation of K is less than some $\lambda < \infty$.

Theorem 1. *If the assumptions (A1)–(A9) hold, for $x \in I, t \in [a, b]$ and if the bandwidth $h = (h_n)$ satisfies $h \rightarrow 0, \ln n/nh \rightarrow 0$ and $nh^5/\ln n = O(1)$, then*

$$\hat{\Lambda}_h^c(t|x) - \Lambda(t|x) = \sum_{i=1}^n \tilde{B}_{hi}(x) \zeta(T_i, \delta_i, \xi_i, \xi_i \nu_i, t, x) + R_n(t, x)$$

with

$$\begin{aligned} \zeta(T_i, \delta_i, \xi_i, \xi_i \nu_i, t, x) &= \frac{\mathbf{1}(T_i \leq t, \delta_i = 1)}{J(T_i^-|x)} \\ &\quad - \int_0^t (\mathbf{1}(T_i \geq v) + \mathbf{1}(T_i < v, \xi_i \nu_i = 1)) \frac{dH^1(v|x)}{J^2(v^-|x)}, \end{aligned} \quad (6)$$

$$\tilde{B}_{hi}(x) = \frac{1}{m(x)} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \quad (7)$$

where $m(x)$ is the density function of the covariate X , $H^1(t|x)$ and $J(t|x)$ are given in (4) and (5), and the remainder term is

$$\sup_{a \leq t \leq b, x \in I} |R_n(t, x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{ a.s.}$$

Theorem 2. *If the assumptions (A1)–(A9) hold, for $x \in I, t \in [a, b]$ and if the bandwidth $h = (h_n)$ satisfies $h \rightarrow 0, \ln n/nh \rightarrow 0$ and $nh^5/\ln n = O(1)$, then*

$$\hat{F}_h^c(t|x) - F(t|x) = \sum_{i=1}^n \tilde{B}_{hi}(x) (1 - F(t|x)) \zeta(T_i, \delta_i, \xi_i, \xi_i \nu_i, t, x) + R_n(t, x)$$

where ζ is defined in (6), $\tilde{B}_{hi}$ in (7) and

$$\sup_{a \leq t \leq b, x \in I} |R_n(t, x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{ a.s.}$$

In Corollary 1 we show that the proposed estimators $\hat{\Lambda}_h^c(t|x)$ and $\hat{S}_h^c(t|x)$ are strongly consistent.

Corollary 1. *If the assumptions (A1)–(A9) hold, for $x \in I, t \in [a, b]$ and if the bandwidth $h = (h_n)$ satisfies $h \rightarrow 0, \ln n/nh \rightarrow 0$ and $nh^5/\ln n = O(1)$, then*

$$\sup_{a \leq t \leq b, x \in I} \left| \hat{\Lambda}_h^c(t|x) - \Lambda(t|x) \right| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) \text{ a.s.}$$

and

$$\sup_{a \leq t \leq b, x \in I} \left| \hat{S}_h^c(t|x) - S(t|x) \right| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) \text{ a.s.}$$

Next proposition provides asymptotic expressions of the bias and the variance of $\hat{S}_h^c(t|x)$.

Proposition 3. *The bias and variance of the nonparametric estimator of the survival function $\hat{S}_h^c(t|x)$ are:*

$$\mu_{h,c}(t, x) = h^2 B_c(t, x) + O(h^4) \quad \text{and} \quad \sigma_{h,c}^2(t, x) = \frac{1}{nh} s_c^2(t, x) + O\left(\frac{h}{n}\right),$$

with

$$B_c(t, x) = \frac{1}{2} (1 - F(t|x)) \frac{2\Phi'_c(x, t, x) m'(x) + \Phi''_c(x, t, x) m(x)}{m(x)} d_K, \quad (8)$$

$$s_c^2(t, x) = \frac{(1 - F(t|x))^2}{m(x)} \Phi_1^c(x, t, x) c_K, \quad (9)$$

where $d_K = \int v^2 K(v) dv$, $c_K = \int K^2(v) dv$, $m(x)$ is the density function of the covariate X , and

$$\Phi_c(y, t, x) = E[\zeta(T, \delta, \xi, \xi\nu, t, x) | X = y] \quad \text{and} \quad \Phi_1^c(y, t, x) = E[\zeta^2(T, \delta, \xi, \xi\nu, t, x) | X = y] \quad (10)$$

with ζ given in (6).

Straightforward calculations show that the functions $\Phi_c(y, t, x)$ and $\Phi_1^c(x, t, x)$ in (10) verify

$$\begin{aligned} \Phi_c(y, t, x) &= \int_0^t \frac{dH^1(v|y)}{1 - H(v^-|x) + H^{11}(v^-|x)} - \int_0^t \frac{1 - H(v^-|y)}{(1 - H(v^-|x) + H^{11}(v^-|x))^2} dH^1(v|x), \\ \Phi_1^c(x, t, x) &= \int_0^t \frac{dH^1(v|x)}{(1 - H(v^-|x) + H^{11}(v^-|x))^2}. \end{aligned}$$

Specifically, the derivatives of $\Phi_c(y, t, x)$ in (8) are

$$\Phi_c'(x, t, x) = -\frac{S'(t|x)}{S(t|x)}$$

and

$$\Phi_c''(x, t, x) = 2 \int_0^t \frac{G_c'(v^-|x)}{1 - G_c(v^-|x)} d\left(\frac{S'(v|x)}{S(v|x)}\right) dv - \frac{S''(t|x)}{S(t|x)},$$

with

$$1 - G_c(t|x) = (1 - G(t|x)) + \pi_1(t, x)(1 - p(x))G_1(t|x)$$

where $\pi_1(t, x) = P(\xi = 1 | \nu = 1, C \leq t, X = x)$ and $G_1(t|x) = P(C \leq t | \nu = 1, X = x)$.

Next theorem shows the asymptotic normality of the proposed estimator of the distribution function.

Theorem 3. Under assumptions (A1)-(A9), then for $x \in I$ and for $t \in [a, b]$ as $n \rightarrow \infty$, it follows that

a) If $nh^5 \rightarrow 0$ and $\frac{(\ln n)^3}{nh} \rightarrow 0$, then

$$\sqrt{nh} \left(\hat{F}_h^c(t|x) - F(t|x) \right) \xrightarrow{d} N(0, s_c^2(t, x))$$

b) If $h = Cn^{-1/5}$ then

$$\sqrt{nh} \left(\hat{F}_h^c(t|x) - F(t|x) \right) \xrightarrow{d} N(C^{5/2} B_c(t, x), s_c^2(t, x))$$

with $B_c(t, x)$ and $s_c^2(t, x)$ given in (8) and (9) respectively.

4. BANDWIDTH SELECTION

From Proposition 3, the asymptotic mean integrated squared error (AMISE) of the estimator $\hat{F}_h^c(t|x)$ is obtained:

$$AMISE(\hat{F}_h^c(t|x)) = \int \left(h^4 B_c^2(v, x) + \frac{1}{nh} s_c^2(v, x) \right) dv, \quad (11)$$

with $B_c(t, x)$ and $s_c^2(t, x)$ are given in (8) and (9) respectively. The bandwidth h minimizing (11) is

$$h_{AMISE}(x) = \left(\frac{\int s_c^2(v, x) dv}{4 \int B_c^2(v, x) dv} \right)^{1/5} n^{-1/5}. \quad (12)$$

Estimating the unknown functions in (12) is a cumbersome task. One possibility to overcome this difficulty is to consider a bootstrap estimation procedure.

The bootstrap bandwidth selector is the bandwidth h minimizing the bootstrap version of the MISE, that can be approximated as follows

$$MISE_x^*(h) \simeq \frac{1}{B} \sum_{b=1}^B \int (\hat{S}_h^{c,*b}(v|x) - \hat{S}_{g_x}^c(v|x))^2 dv, \quad (13)$$

where $\hat{S}_h^{c,*b}(v|x)$ is the nonparametric survival estimator computed with a bandwidth h and the b -th bootstrap resample, and $\hat{S}_{g_x}^c(v|x)$ is the same estimator computed with a pilot bandwidth g_x and the original sample. Considering a bandwidth search grid $\{h_1, \dots, h_L\}$, the procedure for obtaining the bootstrap bandwidth selector of h for a fixed covariate value x , is as follows:

1. Generate B bootstrap resamples of the form $\{(X_i^{(b)}, T_i^{*(b)}, \delta_i^{*(b)}, \xi_i^{*(b)}, \xi_i^{*(b)} \nu_i^{*(b)})\}, i = 1, \dots, n\}$, for $b = 1, \dots, B$.
2. For the b -th bootstrap resample, $b = 1, \dots, B$, compute the nonparametric estimator $\hat{S}_{h_l}^{c,*b}(x)$ with bandwidth $h_l \in \{h_1, \dots, h_L\}$.
3. With the original sample and the pilot bandwidth g_x , compute $\hat{S}_{g_x}^c(x)$.
4. For each bandwidth $h_l \in \{h_1, \dots, h_L\}$, compute the Monte Carlo approximation of $MISE_x^*(h_l)$, given by (13).
5. The bootstrap bandwidth, h_x^* , is the minimizer of $MISE_x^*(h)$ in (13) over the grid of bandwidths $\{h_1, \dots, h_L\}$.

In step 1, we consider the simple weighted bootstrap, without resampling the covariate X , which is equivalent to the simple weighted bootstrap proposed by Li and Datta (2001), and used for bandwidth selection without known cures by López-Cheda et al (2017a). For fixed x and $i = 1, \dots, n$, we set $X_i^* = X_i$ and generate $(T_i^*, \delta_i^*, \xi_i^*, \xi_i^* \nu_i^*)$ from the weighted empirical distribution:

$$\hat{F}_{g_x}(u, v, w, y|x) = \sum_{i=1}^n B_{g_x i}(x) \mathbf{1}(T_i \leq u, \delta_i \leq v, \xi_i \leq w, \xi_i \nu_i \leq y)$$

and $B_{g_x i}(x)$ is the Nadaraya-Watson weight in (1) with pilot bandwidth g_x . The resulting bootstrap resample is $\{(X_i, T_i^*, \delta_i^*, \xi_i^*, \xi_i^* \nu_i^*), i = 1, \dots, n\}$.

Regarding the pilot bandwidth g_x needed to generate the bootstrap resample and in step 3, it is typically chosen asymptotically of order larger than h_x . This oversmoothing pilot bandwidth is required for the bootstrap bias and variance to be asymptotically efficient estimators for the bias and variance terms. The order $n^{-1/9}$ satisfies the conditions for this asymptotically optimal pilot bandwidth in Theorem 1 of Li and Datta (2001), and it coincides with the order obtained by Cao and González-Manteiga (1993) for the uncensored case.

Simulation results in López-Cheda et al (2017a, 2017b) showed that the choice of the pilot bandwidth has a small effect on the final bootstrap bandwidth. Consequently, we propose to use the same local pilot bandwidth as in López-Cheda et al (2017a, 2017b):

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ is the distance from x to the k -th nearest neighbor on the right, $d_k^-(x)$ the distance from x to the k -th nearest neighbor on the left, and k a suitable integer depending on the sample size. If there are not at least k neighbors on the right (or left), we use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$) respectively. Our numerical experience shows that a good choice is to consider $k = n/4$. Note that when $n = 100$ the value of the local pilot bandwidth g_x is the mean distance to the 25th nearest neighbor on both the left and right sides.

5. SIMULATION STUDY

In this section we show the practical performance of the proposed estimator through a simulation study. We considered the survival function

$$S(t|x) = (1 - p(x)) + p(x)S_0(t|x)$$

with latency function

$$S_0(t|x) = \begin{cases} \frac{\exp(-\alpha(x)t) - \exp(-\alpha(x)4.605)}{1 - \exp(-\alpha(x)4.605)} & 0 \leq t \leq 4.605 \\ 0 & t > 4.605 \end{cases} \quad \text{where } \alpha(x) = \exp\left(\frac{x+20}{40}\right).$$

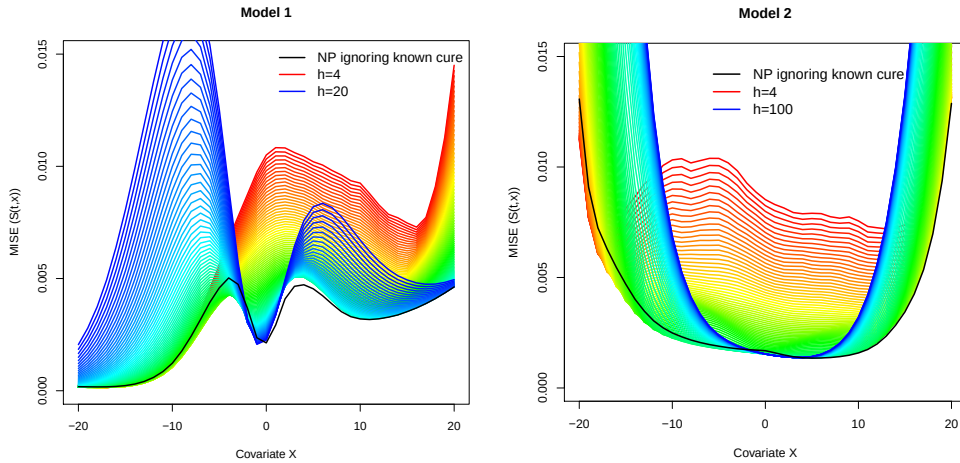


Figure 1: MISE of the proposed nonparametric estimator of the survival function $\hat{S}_h^c(t|x)$ in a grid of 100 bandwidths, from lowest in red ($h_1 = 4$) to highest in blue ($h_{100} = 20$), in Model 1 (left) and Model 2 (right). The MISE of the generalized product-limit estimator in Beran (1981) computed with the optimal bandwidth is given as reference (black line).

We simulated two models according to the cure rate functions:

$$1 - p_1(x) = 1 - \frac{\exp(0.476 + 0.358x)}{1 + \exp(0.476 + 0.358x)} \quad \text{and} \quad 1 - p_2(x) = 0.5 - \frac{1}{16000}x^3$$

for Models 1 and 2, respectively.

The censoring times C were generated from an exponential distribution $\text{Exp}(0.3)$. In this way, the percentage of censoring was 54% with an average cure rate of 47% in Model 1, whereas in Model 2, 61% of the observations were censored and the corresponding cure rate was 50%. The covariate X was generated from a uniform distribution $U[-20, 20]$. In both scenarios, the probability that a cure is observed was $\pi = 80\%$. Data were generated such that the censoring times C and the lifetimes Y are independent conditionally on X . We simulated $m = 1000$ samples of sample size $n = 100$. The number of bootstrap resamples was $B = 1000$.

Figure 1 shows the MISE of the nonparametric estimator of the conditional survival function when cure is partially known, $\hat{S}_h^c(t|x)$, computed in a grid of 100 bandwidths, from $h_1 = 4$ to $h_{100} = 20$, equispaced in a logarithmic scale. The plot includes also the MISE of the generalized product-limit estimator that ignores the known cures (Beran, 1981) computed with the optimal bandwidth. It can be seen that, in Model 2, there are several bandwidths for which the proposed estimator clearly improves or is comparable to the estimator proposed by Beran (1981) with the optimal bandwidth. A similar performance can be seen in Model 1 for most values of the covariate X but for a small range of values between 0 and 8, approximately.

Figure 2 displays the 25th, 50th and 75th quantiles of the simulated bootstrap bandwidths together with the optimal MISE bandwidth. It can be seen how the bootstrap bandwidths approach the optimal ones. In order to assess the impact on the MISE of using the bootstrap bandwidth instead of the optimal one, Figure 3 shows the MISE of the proposed nonparametric estimator of the survival function as a function of the bandwidth h and the covariate value x . For each value of the covariate, the optimal bandwidth is indicated with a red cross. On the other hand, the density of the bootstrap bandwidths is plotted as a grey cloud. It can be seen that the bootstrap bandwidth approximates quite well the optimal one. Besides, vertical lines indicate that for the corresponding value x_0 of X , the MISE of $\hat{S}_h^c(t|x_0)$ tends to be constant as a function of h . Therefore different bandwidths would yield approximately the same MISE. In those cases, the bootstrap bandwidth being far from the optimal bandwidth does not imply a loss of efficiency.

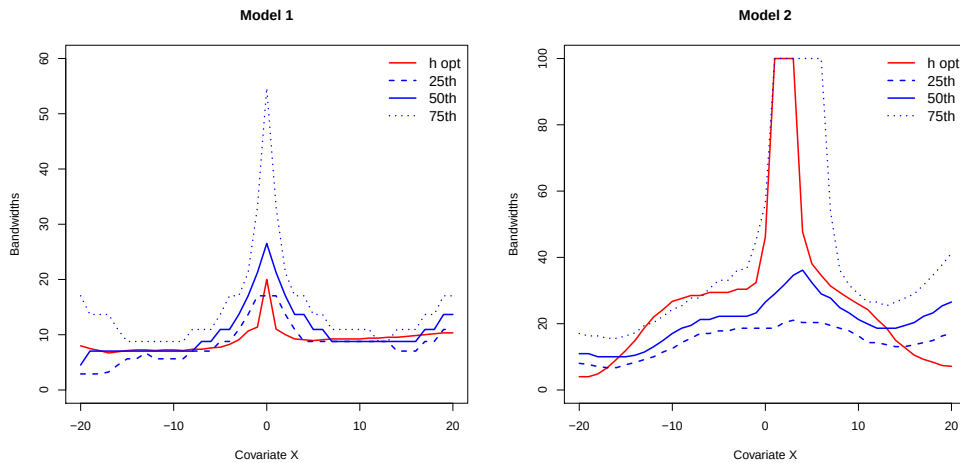


Figure 2: 25th, 50th and 75th quantiles of the simulated bootstrap bandwidths for the nonparametric estimator $\hat{S}_h^c(t|x)$ computed over 1000 trials, in Model 1 (left) and Model 2 (right). The optimal MISE bandwidth is given as reference (red line).

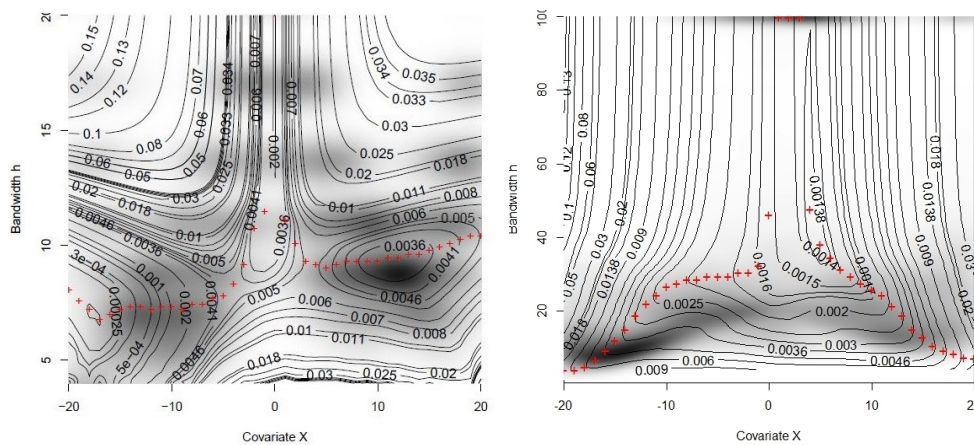


Figure 3: MISE of the proposed nonparametric estimator of the survival function $\hat{S}_h^c(t|x)$ as a function of the bandwidth h and the value of the covariate X , in Model 1 (left) and Model 2 (right). For each value of the covariate, the optimal bandwidth is given by a red cross. The density of the bootstrap bandwidths is shown in grey.

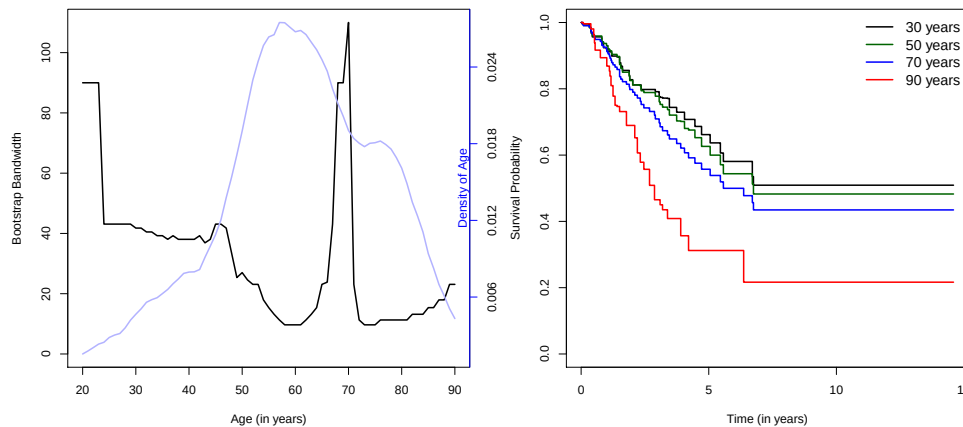


Figure 4: Left: bootstrap bandwidth selector for the estimation of the survival function conditioned on the age; the nonparametric estimator of the density of the age is displayed as reference. Right: proposed nonparametric estimator of the survival function $\hat{S}_h^c(t|x)$ as a function of the covariate age, for 4 different ages.

6. REAL DATA EXAMPLE

In this section we illustrate the proposed estimator using a data set of 233 patients of sarcoma cancer, from the University Hospital of Santiago de Compostela, Spain. The database consists of 261 patients (119 males) aged 20-90 years old. The event of interest is death from sarcoma. Besides clinical covariates such as histological type or tumor depth, each patient is known to be either with tumor or tumor free at the last follow-up. If a patient is tumor free and the survival time exceeds a fixed period of time, set at 5 years in this example, the patient can be diagnosed as cured, that is, he/she is assumed not to die from sarcoma. A total of 195 observations (74.71%) are censored. Among them, a total of 18 patients were tumor free for more than five years. Hence, they were assumed to be long-term survivors.

The age of the patient is used to illustrate the nonparametric estimation of the survival function, when cure is partially known. Figure 4 shows the value of bootstrap bandwidth selector for the different ages. On the right, we see how age effect is relevant to the estimated survival curve. We see that, among all age groups, the survival curves are decreasing with age until they reach a plateau. As it was expected, the survival curve of the 90 years patients is decreasing much faster.

We have only considered age in this analysis for illustration purposes. Future analyses should be extended to other clinical factors to control for confounders.

Acknowledgement

This work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

REFERENCES

- Amico M. and Van Keilegom, I. (2018). Cure models in survival analysis. Review paper, KU Leuven; Leuven, Belgium.
- Beran, R. (1981) Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.
- Bernhardt, P. W. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine* 35, 4607–4623.

- Betensky, R. A. and Schoenfeld, D. A. (2001). Nonparametric estimation in a cure model with random cure times. *Biometrics* 57, 282–286.
- Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)* 11, 15–53.
- Cao, R. and González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics* 2, 379–388.
- Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scandinavian Journal of Statistics* 14, 181–197.
- Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics* 17, 1157–1167.
- Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhyā Series A* 54, 252–259.
- González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Communications in Statistics-Theory and Methods* 4, 65–78.
- Iglesias-Pérez, M. C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics* 10, 213–244.
- Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics* 48, 1223–1234.
- Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Annals of the Institute of Statistical Mathematics* 53, 708–729.
- Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Annals of Statistics* 23, 787–823.
- López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis* 105, 144–165.
- López-Cheda, A., Jácome, M. A. and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *TEST* 26, 353–376.
- Nieto-Baraja, L. E. and Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian Journal of Statistics* 35, 540–556.
- Peng, Y. and Taylor, J. M. G. (2014). Cure models. In: Klein J, Van Houwelingen HC, Ibrahim JG, Scheike TH., eds. *Handbook of Survival Analysis* Boca Raton, FL, USA: Chapman & Hall. pp. 113–134.
- Wu, Y., Lin, Y., Lu, S.E., Li, C.S. and Shih, W. J. (2014). Extension of a Cox proportional hazards cure model when cure information is partially known. *Biostatistics* 15, 540–554.
- Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics* 42, 1–17.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

REXIÓN DE INCIDENCIA DOS INCENDIOS FORESTAIS EN GALICIA

Manuel Antonio Novo Pérez¹, Ana Belén Buide Carballosa¹ e M^a José Ginzo Villamayor²

¹Instituto Tecnolóxico de Matemática Industrial (ITMATI)

²Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela (USC)

RESUMO

Galicia é unha das rexións con maior superficie forestal afectada por incendios forestais. Neste contexto, a correcta xestión dos recursos, tanto aéreos como terrestres é moi importante para permitir afrontar os incendios que poidan xurdir. En concreto, é importante para os equipos de extinción coñecer os lugares onde soen actuar as aeronaves de extinción, tanto durante unha actuación concreta como de forma xeral, así como as rexións con maior incidencia de incendios para coñecer cales son as zonas máis conflictivas e, polo tanto, destinalles maior número de recursos.

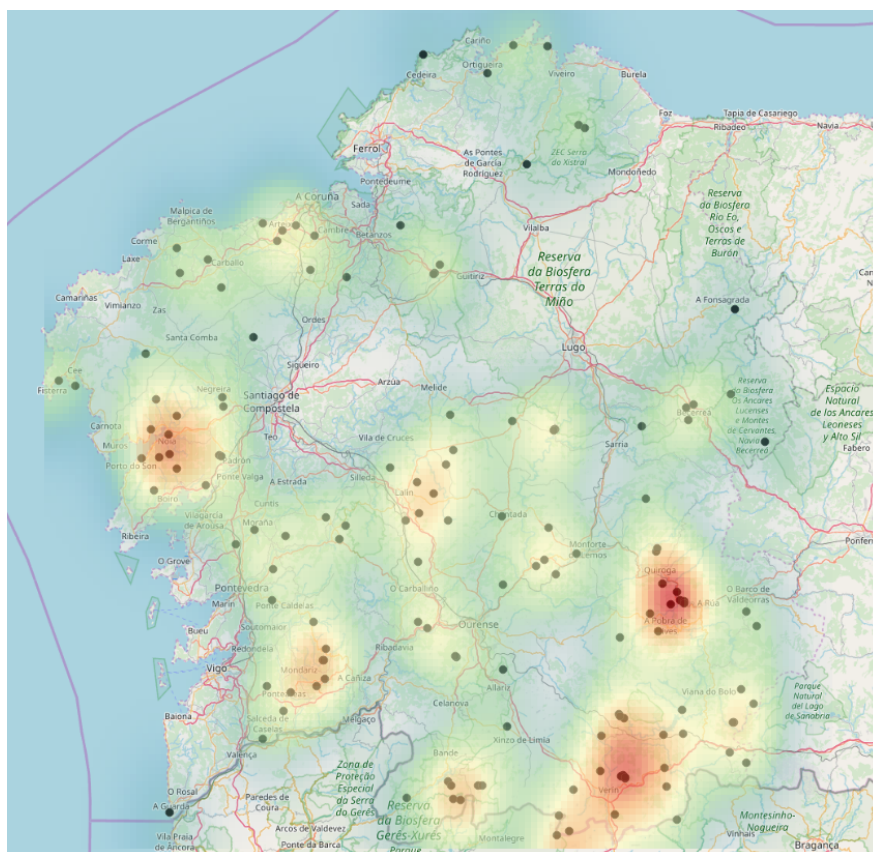


Figura 1: Mapa de incidencia de incendios en Galicia entre marzo e agosto de 2019. En cor negra indícase a localización dos incendios.

Neste traballo amósase como elaborar, utilizando unha técnica estatística como a estimación non paramétrica da función de intensidade, mapas nos que se indican as

rexións de maior incidencia de incendios en Galicia e as principais zonas nas que actúan as aeronaves de extinción, tanto nunha actuación concreta como durante unha campaña de incendios. Tamén se amosa como afecta o problema de selección da ventá e como se pode tratar de resolver. Na Figura 1 pode verse un exemplo de mapa de incidencia de incendios en Galicia entre marzo e agosto do ano 2019.

Para a elaboración deste traballo utilizáronse datos dos voos das aeronaves de extinción durante o ano 2019 en Galicia (procedentes do proxecto Civil UAVs Initiative (CUI), no que participa ITMATI). Estes datos constan dalgúñas variables como o estado da aeronave en cada instante de tempo da súa actuación (lonxitude, latitude, velocidade...) así como algúns eventos, como pode ser o momento no que realizan unha descarga de auga. A partir dos datos infírense os incendios ocorridos nese mesmo período.

Palabras e frases chave: Incendios forestais, intensidade, non paramétrica, aeronaves.

AGRADECEMENTOS

Os investigadores Manuel Novo, Ana Buide e María José Ginzo agradecen o apoio do proxecto CUI da Axencia Galega de Innovación (GAIN) da Xunta de Galicia.

REFERENCIAS

- Baddeley, A., Rubak, E. and Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press.
- Bivand, R., Pebesma, E. and Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

INTERVALOS DE CONFIANZA BOOTSTRAP PARA CURVAS PERCENTÍS DO IMC NA POBOACIÓN INFANTIL GALEGA

María del Carmen Iglesias-Pérez¹, María Arias Vázquez e María Isolina Santiago Pérez²

¹Departamento de Estatística e I.O., Universidade de Vigo

²Departamento de Epidemioloxía da Dirección Xeral de Saúde Pública, Santiago de Compostela

RESUMO

En 2013, a Dirección Xeral de Saúde Pública (DXSP) realizou un estudo antropométrico da poboación escolar galega de 6 a 15 anos, con medidas de 7.438 alumnos representativos desa poboación. Os datos foron usados recentemente para estimar as curvas e valores percentís de peso, altura, índice de masa corporal (IMC), circunferencia de cintura e perímetro de cadeira, en función da idade, en nenos e nenas por separado (Arias, 2019). A estimación realizouse mediante a aplicación dos modelos aditivos xeneralizados de localización, escala e forma (GAMLSS), especificamente co modelo BCPE de Rigby e Stasinopoulos (2004). Neste traballo expónse un método baseado na metodoloxía bootstrap para calcular os intervalos de confianza das curvas percentís obtidas aplicando o modelo BCPE. O procedemento aplícase as curvas percentís do IMC que definen os puntos de corte de baixo peso, sobrepeso e obesidade, segundo tres estudos de referencia no campo da saúde.

Palabras e frases chave: curvas percentís, GAMLSS, intervalos de confianza, bootstrap.

1. INTRODUCCIÓN

As curvas percentís de peso, talla e índice de masa corporal (IMC) na idade infantil e na adolescencia indican como deben crecer os nenos, e permiten aos profesionais sanitarios identificar medidas inusuais que corresponden a valores extremos da distribución, e que poden ser indicativas dun problema de saúde ou de nutrición. En particular, o IMC reflicte o peso relativo á talla para unha idade determinada, e é un indicador do estado ponderal que se emprega para establecer categorías de baixo peso, peso normal, sobrepeso e obesidade.

A distribución das medidas antropométricas vai cambiando de forma continua coa idade, polo que o cálculo de percentís debe facerse mediante a estimación de curvas suavizadas ao longo da idade. Cole propuso en 1990 o método LMS baseado en asumir que a variable das medicións segue unha distribución Box-Cox normal con tres parámetros que dan nome ao método: a potencia da transformación (L), a media (M), e o coeficiente de variación (S). As curvas percentís obtéñense a partir das curvas L, M e S, que se estiman de forma suave en función da idade polo método de máxima verosimilitude penalizada. O método BCPE ou LMSP, introducido por Rigby e Stasinopoulos (2004), é unha extensión do método LMS que asume unha distribución Box-Cox Power Exponential con 4 parámetros.

Existen diversas curvas percentís de referencia, tanto a nivel internacional como nacional, que foron obtidas aplicando os métodos clásicos LMS ou BCPE. Entre as curvas internacionais destacan as publicadas pola Organización Mundial da Saúde (OMS) en 2007, e as da International Obesity Task Force, desenvolvidas por Cole e Green para a clasificación do sobrepeso e a obesidade no ano 2000 e para o baixo peso en 2006. En España, o estudo máis importante de crecemento infantil é o da Fundación Orbegozo, cuxas curvas de percentís do ano 2011 son as que se empregan para monitorar a evolución do crecemento nos controis pediátricos de saúde. Nos tres estudos os valores de referencia para baixo peso, sobrepeso e obesidade obtivéronse calculando os percentís do IMC que aos 18 anos igualaban os puntos de corte de adultos (IMC de 18,5, 25 e 30 para baixo peso,

	OMS		Cole		Orbegozo	
	Nenos	Nenas	Nenos	Nenas	Nenos	Nenas
Baixo peso	P2,3	P2,3	P15,5	P16,5	P4	P10
Sobrepeso	P84,1	P84,1	P90,5	P89,3	P79	P89
Obesidade	P97,7	P97,7	P98,9	P98,6	P97,5	P99

Táboa 1: Percentís de referencia de baixo peso, sobrepeso e obesidade en nenos e nenas nos estudos da OMS, Cole e Orbegozo.

sobrepeso e obesidade, respectivamente). Na Táboa 1 móstrase a orde dos percentís de referencia en nenos e nenas para cada un dos estudos.

No ano 2013, a Dirección Xeral de Saúde Pública realizou un estudo antropométrico da poboación escolar galega de 6 a 15 anos, no que se recolleron medicións de 7.438 alumnos representativos desa poboación, e que foron seleccionados mediante unha mostraxe bietápica estratificada de centros e aulas. Cos datos dese estudio estimáronse as curvas percentís de peso, talla, IMC, perímetro de cintura e perímetro de cadeira, en función da idade, e para nenos e nenas por separado. O método empregado foi o modelo aditivo xeneralizado de localización, escala e forma (GAMLSS), que xeneraliza os métodos clásicos de estimación de curvas percentís, LMS e BCPE, e que está implementado na librería gamlss de R. Dado que os datos empregados proceden dunha enquisa e, polo tanto, están suxeitos ao erro de mostraxe, neste traballo propónse un método bootstrap para obter intervalos de confianza (IC) das curvas percentís. En particular, estimáronse os IC dos percentís do IMC correspondentes aos valores de referencia de baixo peso, sobrepeso e obesidade dos estudos de Orbegozo, OMS e Cole (Táboa 1).

2. METODOLOXÍA

Este apartado describe a metodoloxía empregada no traballo. En primeiro lugar introdúcese a formulación xeral dos modelos GAMLSS, defínese o caso particular do método BCPE de Rigby e Stasinopoulos (2004) empregado na estimación das curvas percentís do IMC por idade, e explícase brevemente o proceso de estimación dos modelos con splines cúbicos. A continuación, preséntase un método de tipo bootstrap para calcular intervalos de confianza das curvas percentís estimadas.

2.1 Os modelos GAMLSS

Os modelos GAMLSS asumen unha distribución paramétrica para a variable resposta Y e permiten o modelado, non só do parámetro de localización (a media ou a mediana), senón tamén dos parámetros de dispersión e forma, como funcións dunha ou varias variables explicativas.

Matemáticamente, sexa $\mathbf{y}^T = (y_1, \dots, y_n)$ o vector n -dimensional da variable resposta, onde $y_i \sim F(y_i|\boldsymbol{\theta}^i)$, sendo $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ un vector de parámetros. Os parámetros μ_i e σ_i con $i = 1, 2, \dots, n$ son parámetros de localización e de escala (media ou mediana, e coeficiente de variación ou varianza, respectivamente), e ν_i e τ_i con $i = 1, 2, \dots, n$ son parámetros de forma (asimetría e curtose). Sexan $g_k(\cdot)$ funcións de enlace coñecidas, para $k = 1, \dots, p$, sendo p o número de parámetros. Na maioría dos casos prácticos, hai como máximo catro parámetros ($p = 4$). A relación entre os parámetros da distribución e as variables explicativas ven dada por:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk}, \quad (1)$$

onde $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ son vectores n -dimensionais, $\boldsymbol{\beta}_k$ un vector de $\mathbf{J}'_k$ parámetros, $\boldsymbol{\gamma}_{jk}$ variable aleatoria q_{jk} -dimensional, $\mathbf{X}_k$ e $\mathbf{Z}_{jk}$ matrices de deseño fixo de dimensión $n \times \mathbf{J}'_k$ e $n \times q_{jk}$, respectivamente.

Esta definición combina termos paramétricos ($\mathbf{X}_k \boldsymbol{\beta}_k$) e termos aditivos ($\mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk}$) no denominado predictor lineal, $\boldsymbol{\eta}_k$, permitindo unha formulación xeral de moitos modelos ou submodelos. En concreto, os termos aditivos poden modelizar o suavizado non paramétrico adoptando a seguinte

forma particular:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \quad (2)$$

onde h_{jk} é unha función de suavización da covariable X_{jk} que hai que estimar. Xa que $\boldsymbol{\gamma}$ representa un efecto aleatorio, a suavización non paramétrica estase a escribir como un modelo de efectos aleatorios.

Estes modelos son estimados por maximización do logaritmo da función de verosimilitude penalizada, dada pola seguinte expresión xeral: $\ell_p = \ell - \frac{1}{2} \sum_{k=1}^p \sum_{j=1}^{J_k} \lambda_{jk} \boldsymbol{\gamma}_{jk}^t \mathbf{G}_{jk} \boldsymbol{\gamma}_{jk}$, onde ℓ é o logaritmo da función de verosimilitude, λ_{jk} son parámetros de suavización chamados hiperparámetros e $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ é unha matriz simétrica de dimensión $q_{jk} \times q_{jk}$, cuxa expresión depende do método de suavización que se usa. Neste caso utilizáronse splines cúbicos.

De modo que é necesario estimar $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ para axustar o modelo. Para isto, utilízase o algoritmo RS que maximiza á verosimilitude penalizada con respecto a $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ para $\boldsymbol{\lambda}$ dado. Os parámetros $\boldsymbol{\lambda}$ foron estimados minimizando o GAIC, onde $GAIC(\#) = -2\ell + \#df$, df denota os graos efectivos de liberdade empregados no modelo (que representan a dimensión da parte paramétrica e dos efectos aleatorios e teñen relación cos hiperparámetros) e $\#$ denota unha penalización fixa.

2.2 Os métodos LMS e BCPE

Para distribucións unimodais e asimétricas, Cole definiu unha transformación de Y dependente de 3 parámetros (L, M e S) para obter normalidade, o que proporciona unha fórmula sinxela dos percentís. O método LMS modela cada un dos tres parámetros como unha función suave (spline) da idade. Non obstante, o método de Cole non corrixe a presenza de curtose, polo que se introduciu o método BCPE (ou LMSP), que xeneraliza o modelo de Cole introducindo un cuarto parámetro adicional (P = potencia) para corrixir a curtose. A transformación de Cole vén dada pola seguinte expresión:

$$z_{LMS} = \begin{cases} \frac{1}{\sigma\nu} \left[\left(\frac{Y}{\mu} \right)^\nu - 1 \right] & \text{se } \nu \neq 0, \\ \frac{1}{\sigma} \log \left(\frac{Y}{\mu} \right) & \text{se } \nu = 0, \end{cases} \quad (3)$$

para $0 < Y < \infty$, con $\mu, \sigma > 0$. Ademais, μ é o parámetro M que representa a media ou mediana, σ é o parámetro S que representa o coeficiente de variación e ν é o parámetro L que representa a asimetría. De aquí proceden as siglas L, M e S do método de Cole.

Unha variable Y que a través da transformación (3) segue unha distribución normal estándar, denomínase distribución Box-Cox Normal ou Distribución Box-Cox Cole e Green (BCCG). Se en vez dunha distribución normal estándar, z_{LMS} en (3) segue unha distribución power exponencial estándar de parámetro $\tau > 0$, (parámetro de potencia P que da nome a o método LMSP), entón dise que Y segue unha distribución Box Cox Power Exponencial (BCPE).

O método LMS supón que Y ten unha distribución BCCG, con tres parámetros (μ, σ, ν) , mentres que o método BCPE (ou LMSP) asume unha distribución BCPE, con catro parámetros (μ, σ, ν, τ) . O método BCPE xeneraliza ao método LMS, xa que cando $\tau = 2$ a distribución BCPE redúcese a BCCG.

Polo tanto, o modelo BCPE é un caso particular dun modelo GAMLSS cunha única variable explicativa $X = Idade$, que se formula da forma seguinte:

Dado $X = x$, Y segue unha distribución $BCPE(\mu, \sigma, \nu, \tau)$ onde cada un dos parámetros é modelado como unha función non paramétrica suave de x , é dicir, $g_1(\mu) = h_1(x)$, $g_2(\sigma) = h_2(x)$, $g_3(\nu) = h_3(x)$, $g_4(\tau) = h_4(x)$, sendo $g_k(\cdot)$ a función identidade para $k = 1, 3$, e o logaritmo para $k = 2, 4$.

Unha vez estimados os parámetros da distribución (μ, σ, ν, τ) en función dos valores de X , mediante a maximización da verosimilitude penalizada dun modelo GAMLSS, pódese calcular o percentil de orden α mediante a expresión:

$$Y_\alpha = \begin{cases} \mu [1 + \sigma\nu z_\alpha]^\frac{1}{\nu} & \text{se } \nu \neq 0, \\ \mu \exp(\sigma z_\alpha) & \text{se } \nu = 0, \end{cases} \quad (4)$$

onde z_α é o percentil de orden α dunha distribución power exponencial estándar con parámetro de potencia τ .

2.3 Estimación dos modelos co paquete `gamlss` de R

A estimación das curvas percentis do IMC en nenos e nenas de 6 a 15 anos en Galicia realizouse co paquete `gamlss` de R.

O primeiro paso foi seleccionar unha distribución adecuada para Y (IMC), axustando modelos con diferentes familias de distribucións (normal, BCCG e BCPE) para cada sexo con p-splines, utilizando as funcións `gamlss` e `pb()` de R. A función `pb()` ten a vantaxe de que calcula os hiperparámetros de forma automática. Entre os modelos axustados elixiuse o de menor valor GAIC (2,5) que resultou ter distribución BCPE.

Unha vez seleccionada a distribución, procedeuse a suavizar os termos aditivos dos modelos mediante splines cúbicos, a través da función `cs()`. Neste caso, foi necesaria a estimación previa dos hiperparámetros utilizando a función `find.hyper`. Esta función requiriu: valores iniciais baixos, medios e altos para asegurar a converxencia a un mínimo global de GAIC (#) e diferentes penalizacións (neste caso $\# = 2, 2.5, 3, \log(n)$) para atopar o modelo máis adecuado. Os hiperparámetros estimados usáronse para axustar diferentes modelos para cada penalización.

O seguinte paso foi a validación e comparación dos modelos axustados a través das funcións `worm.plot` e `Q.stats`. Estas funcións comprobaban que os modelos se axustan ben aos datos e estudan a normalidade dos residuos. O modelo que mellor se adapta, para cada variable, foi seleccionado baixo a notación, BCPE($df_\mu, df_\sigma, df_\nu, df_\tau$). No caso do IMC, para os nenos o modelo seleccionado foi BCPE(6,4,3,4) e para as nenas foi BCPE(6,4,3,6). (Ver Arias, 2019).

Finalmente, as funcións `centiles` e `centiles.pred` permitiron debuxar as curvas percentís do modelo seleccionado e calcular valores percentís para novos puntos x .

2.4 Cálculo dos intervalos de confianza bootstrap

Para o cálculo dos intervalos de confianza das curvas percentís empregouse unha metodoloxía tipo bootstrap que se describe a continuación.

Sexan, $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ a mostra observada, BCPE($\hat{df}_\mu, \hat{df}_\sigma, \hat{df}_\nu, \hat{df}_\tau$) o modelo estimado mediante a metodoloxía descrita no apartado anterior, e $\hat{Y}_\alpha(x)$ a estimación do percentil de orden α no punto x , onde α e x son valores fixados previamente. (Considerando un vector de valores de x obtense a curva percentil estimada). Os pasos do método son os seguintes:

1. Escoller o número de remostras bootstrap B . Neste traballo $B = 250$.
2. Calcular B mostras de tamaño n obtidas por mostraxe aleatoria con reempazo da mostra inicial. Cada unha de esas mostras denótase $\{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \dots, (X_n^*, Y_n^*)\}^b$, con $b = 1, \dots, B$.
3. Para cada unha das B mostras bootstrap, axustar un modelo BCPE con splines cúbicos e graos de liberdade estimados coa mostra inicial, isto é $(\hat{df}_\mu, \hat{df}_\sigma, \hat{df}_\nu, \hat{df}_\tau)$; con ese modelo estimado facer a estimación do percentil α no punto x , $\hat{Y}_\alpha^*(x)$. Repetir para as B mostras.
4. Ordenar os B valores $\hat{Y}_\alpha^*(x)$, e calcular os cuantís correspondentes as probabilidades 0,025 e 0,975 que proporcionan o intervalo de confianza ao 95%.

Nota: Se en cada paso se considera un vector de valores x en lugar dun único valor, pódese obter unha serie de intervalos de confianza puntuais, cada un con nivel de confianza do 95%.

O método bootstrap proposto segue un enfoque non paramétrico, xa que no paso 2 cada remostra obtense mediante mostraxe aleatoria con reempazo dos datos orixinais. Supoñendo que o modelo BCPE é correcto, un método bootstrap de tipo paramétrico podería ser máis eficiente. No enfoque paramétrico, o paso 2 do método modifícase de tal xeito que as remostras se obteñen a partir do modelo BCPE axustado coa mostra orixinal.

3. RESULTADOS

Neste apartado móstrase un estudo comparativo das curvas percentís do IMC obtidas cos datos do estudo da DXSP, xunto cos seus intervalos de confianza bootstrap (método non paramétrico), e das curvas obtidas nos estudos de Orbegozo, da OMS e de Cole (Figuras 1, 2 e 3, respectivamente), para nenos e nenas por separado. Finalmente, a Figura 4 mostra unha comparación entre os intervalos de confianza bootstrap calculados aplicando o método paramétrico e o método non paramétrico, correspondente aos percentís de referencia do estudo nacional de Orbegozo.

Os intervalos de confianza (Figuras 1, 2 e 3) non son moi amplos indicando que as curvas percentís estimadas coa enquisa da DXSP representan razoablemente os percentís da poboación infantil galega. Nas tres figuras pódese observar que a amplitude dos intervalos aumenta a medida que a orde do percentil crece e que a amplitude é maior nas idades máis altas.

Respecto ao estudo nacional de Orbegozo (Figura 1) obsérvase que os percentís correspondentes ao baixo peso en Galicia son similares en ambos estudos tanto para nenos como para nenas. Non obstante, ao comparar os percentís que definen o exceso de peso e a obesidade, obsérvase que os valores de IMC obtidos no estudo de Galicia son maiores, especialmente nas nenas.

As diferenzas entre os valores do IMC obtidos no estudo de Galicia e os estudos internacionais da OMS e Cole (Figuras 2 e 3) son moito maiores.

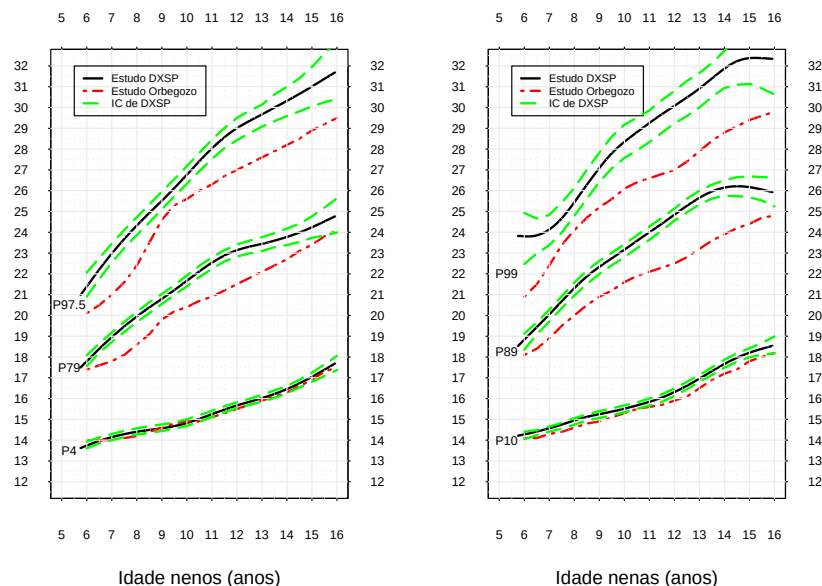


Figura 1: Comparación das curvas percentís do IMC que definen baixo peso, sobrepeso e obesidade cos datos do estudo de Orbegozo (vermello) e cos datos do estudo da DXSP (negro) xunto cos intervalos de confianza correspondentes ao último estudo (verde), para nenos (esquerda) e nenas (dereita), respectivamente.

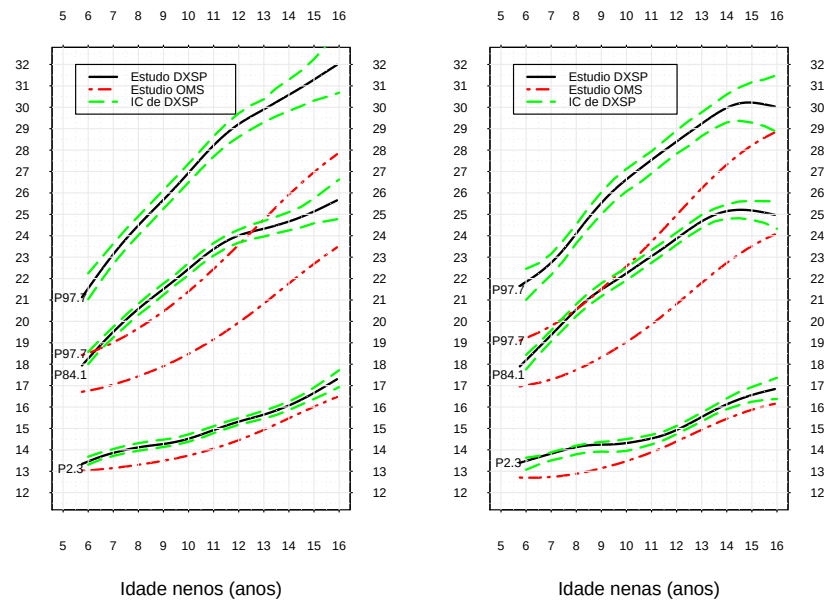


Figura 2: Comparación das curvas percentís do IMC que definen baixo peso, sobrepeso e obesidade cos datos da OMS (vermello) e cos datos do estudo da DXSP (negro) xunto cos intervalos de confianza correspondentes ao último estudo (verde), para nenos (esquerda) e nenas (dereita), respectivamente.

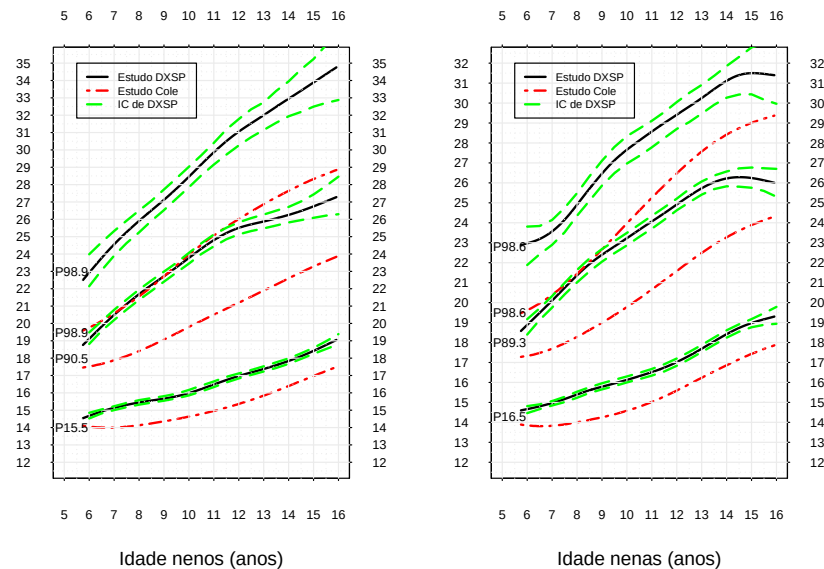


Figura 3: Comparación das curvas percentís do IMC que definen baixo peso, sobrepeso e obesidade cos datos do estudo de Cole (vermello) e cos datos do estudo da DXSP (negro) xunto cos intervalos de confianza correspondentes ao último estudo (verde), para nenos (esquerda) e nenas (dereita), respectivamente.

Os intervalos de confianza calculados polo método paramétrico e os obtidos polo método non paramétrico para os percentís de referencia do estudo Orbegozo (Figura 4) non presentan diferenzas relevantes.

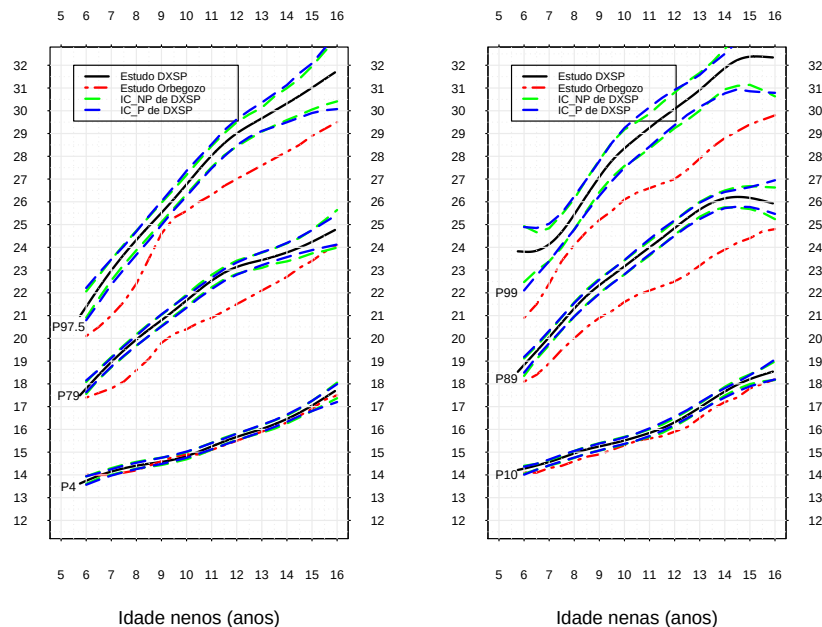


Figura 4: Comparación das curvas percentís do IMC que definen baixo peso, sobrepeso e obesidade cos datos do estudo de Orbegozo (vermello) e cos datos do estudo da DXSP (negro) xunto cos intervalos de confianza para DXSP obtidos mediante bootstrap paramétrico (azul) e bootstrap non paramétrico (verde), para nenos (esquerda) e nenas (dereita), respectivamente.

4. CONCLUSIÓNS

Os intervalos de confianza bootstrap das curvas percentís do IMC obtidos neste traballo non son moi amplos, indicando que as curvas percentís estimadas coa enquisa da DXSP representan razoablemente os percentís da poboación infantil galega.

A amplitude reducida dos intervalos de confianza tamén fai máis confiables os resultados da comparación dos percentís que definen baixo peso, sobrepeso e obesidade nos estudos de Orbegozo, OMS e Cole e os percentís cos datos do estudo da DXSP.

AGRADECEMENTOS

A primeira autora quere agradecer a axuda do Profesor Mikis Stasinopoulos no deseño do método bootstrap aplicado neste traballo. Agradece tamén o apoio do proxecto MTM2017-89422-P (MINECO / AEI / FEDER, UE).

REFERENCIAS

- Arias Vázquez, M. (2019) Curvas de crecimiento de la población infantil gallega. Trabajo Fin de Máster. Máster en Técnicas Estadísticas. Universidad de Santiago de Compostela.
- Cole T. J. and Green P. J. (1992) Smoothing reference centile curves: The lms method and penalized likelihood. *Stat. Med.*, 10, 13051319.
- Cole T. J., Bellizzi M. C., Flegal K. M. and Dietz W. H. (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ*, 320, 124045.

- Cole T. J., Flegal K. M., Nicholls D. and Jackson A. A. (2007) Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ*. 335(7612), 194.
- Fernández, C., Lorenzo, H., Vrotsou, K., Aresti, U., Rica, I., and Sánchez, E. (2011) Estudio de crecimiento de Bilbao. Curvas y tablas de crecimiento (estudio transversal). Bilbao: Fundación Faustino Orbegozo Eizaguirre.
- Rigby, R.A. and Stasinopoulos D.M. (2004) Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution *Statist. Med.*, 23, 3053–3076.
- Rigby, R.A. and Stasinopoulos D.M. (2005) Generalized additive models for location, scale and shape. *Appl.Statist.*, 54, Part 3, 507–554.
- Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017) Flexible regression and smoothing: using GAMLSS in R. Chapman and Hall/CRC
- WHO Multicentre Growth Reference Study Group, and de Onis, M. (2006) WHO Child Growth Standards based on length/height, weight and age. *Acta paediatrica*, 95, 76–85.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

UNA REVISIÓN DE LOS CONTRASTES DE EQUIVALENCIA DE LAS MEDIAS DE DOS POBLACIONES

Javier Rey Ramírez¹, Juan Carlos Pardo Fernández²

¹Grupo SiDOR, Universidade de Vigo

²Departamento de Estatística e Investigación Operativa, Universidade de Vigo

RESUMEN

En la teoría clásica, la comparación de las medias de dos poblaciones se realiza mediante un contraste en el cual la hipótesis nula establece que las medias de ambas poblaciones son iguales frente a la hipótesis alternativa de que son distintas. Sin embargo, cuando el interés práctico es verificar la igualdad de las medias, este tipo de contrastes presentan un inconveniente, ya que cuando el tamaño muestral es elevado la potencia aumenta y el contraste es capaz de detectar diferencias muy pequeñas entre las medias que en la práctica pueden resultar no relevantes.

Surgen entonces los tests de equivalencia, en los cuales la hipótesis nula establece que las medias distan más de una cierta cantidad pre-especificada frente a la alternativa de que la diferencia entre las medias es menor que esa cantidad. Más concretamente, sean X e Y dos variables aleatorias con medias μ_X y μ_Y , respectivamente, y sea $d = \mu_X - \mu_Y$ la diferencia entre las medias. Dadas dos cantidades prefijadas d_L y d_U cumpliendo que $d_L < d_U$, el test de equivalencia contrasta la hipótesis nula $H_0 : d \notin [d_L, d_U]$ frente a la hipótesis alternativa $H_1 : d \in (d_L, d_U)$. Con este planteamiento, el rechazo de la hipótesis nula permitiría concluir que la diferencia de medias se encuentra en el intervalo prefijado (d_L, d_U) . De gran interés práctico resulta el caso $d_U = -d_L = d$, donde $d > 0$ es una cantidad pequeña, ya que un resultado significativo del test permitiría justificar que las medias distan menos de d . En este caso, podemos decir que las medias son equivalentes.

En este trabajo se revisarán los principales tests de equivalencia usados en la práctica. Véanse, por ejemplo, el *two one-sided test* desarrollado por Schuurmann (1987), el test de Anderson y Hachuk (1984) o el test uniformemente más potente invariante de Wellek (2010), que están inspirados en las ideas del t-test clásico, o la adaptación del test de Wilcoxon-Mann-Whitney propuesta en Wellek (1996). Se describirán los distintos procedimientos y se explicarán sus condiciones de aplicabilidad. Además se desarrollará un nuevo test basado en bootstrap. Se comparará el funcionamiento práctico de estos métodos mediante un estudio de simulación y se mostrará una aplicación a datos reales.

Palabras clave: Bootstrap, contraste de equivalencia, medias.

REFERENCIAS

- Anderson, S., Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, 12, 2663-2692.
- Schuurmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.
- Wellek, S. (1996). A new approach to equivalence assessment in standard comparative bioavailability trials by means of the Mann-Whitney statistic. *Biometrical Journal*, 38, 695-710.
- Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

EFECTO DE LA SUAVIZACIÓN EN CONTRASTES BASADOS EN LOS RESIDUOS DE LA REGRESIÓN

Natalia Pérez Veiga¹, Juan Carlos Pardo Fernández²

¹Grupo SiDOR, Universidade de Vigo.

²Departamento de Estadística e IO, Universidade de Vigo.

RESUMEN

En el contexto de la regresión se han diseñado múltiples procedimientos de contraste basados en la distribución de los residuos. Habitualmente, las funciones de distribución involucradas en los estadísticos de contraste se estiman usando la función de distribución empírica, que, por su carácter discontinuo, puede dar lugar a estimaciones pobres de la verdadera función de distribución. En este estudio se comprueba que reemplazando la empírica por su versión suavizada se obtiene una mejora en la potencia de los contrastes basados en la distribución de los residuos para los casos particulares de contraste de bondad de ajuste y comparación de curvas de regresión.

Palabras y frases clave: Bootstrap, Bondad de ajuste, Comparación de curvas de regresión, Regresión no paramétrica, Suavización tipo núcleo

1. INTRODUCCIÓN

Sea (X, Y) un vector aleatorio, donde Y es la variable de interés (o una transformación de la misma) y X una covariable y sea (X_i, Y_i) , $i = 1, \dots, n$ una muestra aleatoria simple de (X, Y) . Supongamos que X e Y satisfacen el modelo de regresión heterocedástico

$$Y = m(X) + \sigma(X)\varepsilon, \quad (1)$$

donde $m(X) = E(Y|X)$ y $\sigma^2(X) = Var(Y|X)$. A la hora de hacer contrastes es habitual suponer que el error y la covariable son independientes, $\varepsilon \perp X$. Einmahl y Van Keilegom (2008) proponen un método cuyo objetivo es contrastar dicha independencia. Dadas $\hat{m}(x)$ y $\hat{\sigma}^2(x)$ estimaciones tipo núcleo de $m(x)$ y $\sigma^2(x)$, respectivamente, se contrasta la independencia de X_i y los residuos $\hat{\varepsilon}_i = (Y_i - \hat{m}(X_i)) / \hat{\sigma}(X_i)$. Para ello se usa un estadístico de contraste que mide la distancia entre la función de distribución conjunta y el producto de las distribuciones marginales. Así, el estadístico de tipo Kolmogorov-Smirnov toma la forma

$$T_{KS} = \sqrt{n} \sup_{x,t} \left| \hat{F}_{X\hat{\varepsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\varepsilon}}(t) \right|,$$

donde $\hat{F}_{X\hat{\varepsilon}}(x,t)$, $\hat{F}_X(x)$ y $\hat{F}_{\hat{\varepsilon}}(t)$ son las funciones de distribución empíricas.

Recientemente, Racine y Van Keilegom (2019) proponen reemplazar las funciones de distribución empíricas por un estimador suavizado de las funciones de distribución y comprueban que se obtiene una importante mejora en la potencia del contraste. En este trabajo se pretende extender esta idea a los contrastes de bondad de ajuste y a la comparación de curvas de regresión mediante un estudio de simulación.

2. CONTRASTE DE BONDAD DE AJUSTE

Para el contraste de bondad de ajuste se supone el modelo (1) con $\varepsilon \perp X$ y se desea contrastar la hipótesis nula $H_0 : m \in \mathcal{M}$ frente a $H_1 : m \notin \mathcal{M}$, donde $\mathcal{M} = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ es una clase paramétrica de funciones de regresión.

Van Keilegom, González-Manteiga y Sánchez-Sellero (2008) proponen emplear estadísticos de contraste basados en la distancia entre la distribución empírica de los residuos no paramétricos, $\hat{\varepsilon}_i$, y la distribución empírica de los residuos estimados bajo la hipótesis nula $\hat{\varepsilon}_{0i} = (Y_i - \hat{m}_0(X_i)) / \hat{\sigma}(X_i)$, siendo $\hat{m}_0$ un estimador paramétrico suavizado de m .

El estadístico de contraste de tipo Kolmogorov-Smirnov basado en las funciones de distribución empíricas tiene la forma

$$T_{KS} = \sqrt{n} \sup_y |\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon_0}(y)|.$$

Como lo habitual es que ε sea una variable continua, F_ε y F_{ε_0} se pueden estimar mediante un estimador tipo núcleo, es decir

$$\hat{F}_\varepsilon^h(y) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^y \frac{1}{h} k\left(\frac{s - \hat{\varepsilon}_i}{h}\right) ds = \frac{1}{n} \sum_{i=1}^n K\left(\frac{y - \hat{\varepsilon}_i}{h}\right)$$

y

$$\hat{F}_{\varepsilon_0}^{h_0}(y) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^y \frac{1}{h_0} k\left(\frac{s - \hat{\varepsilon}_{0i}}{h_0}\right) ds = \frac{1}{n} \sum_{i=1}^n K\left(\frac{y - \hat{\varepsilon}_{0i}}{h_0}\right),$$

donde h y h_0 son parámetros ventana, k es una función de densidad núcleo univariante y K es su correspondiente función de distribución. Luego, el estadístico de contraste de tipo Kolmogorov-Smirnov basado en la suavización de las distribuciones empíricas resulta

$$T_{KS,s} = \sqrt{n} \sup_y |\hat{F}_\varepsilon^h(y) - \hat{F}_{\varepsilon_0}^{h_0}(y)|.$$

La selección de los parámetros ventana se puede realizar minimizando una función de validación cruzada de la forma

$$CV(h) = \frac{1}{nN_j} \sum_{i=1}^n \sum_{j=1}^{N_j} \left(I(\hat{\varepsilon}_i \leq e_j) - \hat{F}_{\varepsilon, -i}^h(e_j) \right)^2, \quad (2)$$

donde e_j , $j = 1, \dots, N_j$ son los puntos de evaluación y $\hat{F}_{\varepsilon, -i}^h$ es el estimador suavizado de F_ε obtenido tras eliminar el dato i -ésimo de la muestra.

Se ha llevado a cabo un estudio de simulación para comparar el funcionamiento del estadístico de Kolmogorov-Smirnov usando funciones de distribución empíricas y suavizadas. Para aplicar ambos contrastes en la práctica se usa un procedimiento bootstrap suavizado de los residuos estandarizados para aproximar la distribución de los estadísticos de contraste. Sean $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_1$ los residuos $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ estandarizados para tener media 0 y varianza 1 y sea $\tilde{F}_\varepsilon$ su distribución empírica. El esquema de remuestreo es el siguiente, siendo $b = 1, \dots, B$ el índice de cada réplica bootstrap.

1. Sea $\varepsilon_1^*, \dots, \varepsilon_n^*$ una muestra independiente e idénticamente distribuida de la distribución de los residuos estandarizados suavizada, es decir, una muestra de $V + cZ$ donde V tiene distribución $\tilde{F}_\varepsilon$, $Z \sim N(0, 1)$ y c es una constante suavizadora relacionada con el tamaño muestral.
2. Se define la nueva variable respuesta como $Y_i^* = m_{\hat{\theta}}(X_i) + \hat{\sigma}(X_i)\varepsilon_i^*$.
3. A partir de la remuestra bootstrap $\{(X_i, Y_i^*), i = 1, \dots, n\}$ se calcula el valor de los estadísticos de contraste: $T_{b,KS}^*$ y $T_{b,KS,s}^*$.

Si $T_{(b),KS}^*$ y $T_{(b),KS,s}^*$ son los estadísticos ordenados de los valores $T_{b,KS}^*$ y $T_{b,KS,s}^*$ obtenidos en el paso 3, entonces $C_{KS} = T_{([B(1-\alpha)]),KS}^*$ y $C_{KS,s} = T_{([B(1-\alpha)]),KS,s}^*$ aproximan los cuantiles $(1 - \alpha)$ de la distribución de T_{KS} y $T_{KS,s}$ bajo la hipótesis nula, respectivamente. Se rechaza la hipótesis nula con nivel α cuando el valor del estadístico T_{KS} (análogamente, $T_{KS,s}$) es mayor que el cuantil C_{KS} (análogamente, $C_{KS,s}$).

El modelo simulado es $Y_i = \theta X_i + a(X_i) + \frac{1}{5}(1 + X_i)\varepsilon_i$, $i = 1, \dots, n$ donde $X_1, \dots, X_n$ son independientes e uniformemente distribuidas en el intervalo unidad y $\varepsilon_1, \dots, \varepsilon_n$ son variables normales estándar independientes. Se desea contrastar la hipótesis nula de que la función de regresión pertenece al modelo paramétrico $H_0 : m(x) = \theta x$, $\theta \in \mathbb{R}$. El término $a(x) = \delta \sin(4\pi x)$, $\delta \in \{0, 0.1, 0.2, \dots\}$, representa la desviación con respecto a la hipótesis nula. El verdadero valor del parámetro es $\theta_0 = 1$.

En la Figura 1 se representa la proporción de rechazos empírica obtenida para una rejilla de posibles valores de δ . Esta simulación se basa en $M = 1000$ muestras con $B = 500$ réplicas bootstrap para los tamaños de muestra $n = 50$ y $n = 200$ y tres niveles de significación, 10%, 5% y 1%. Se puede ver que hay una mejora en la potencia del contraste en todos los casos, especialmente para $\alpha = 0.01$.

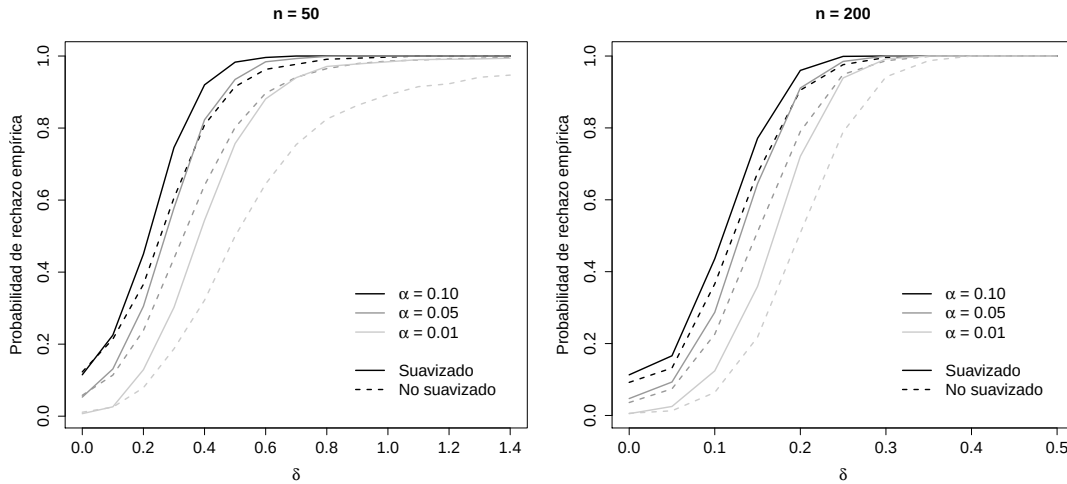


Figura 1: Curvas de potencia para el contraste de bondad de ajuste.

3. CONTRASTE DE COMPARACIÓN DE CURVAS

Sean ahora (X_j, Y_j) , $1 < j < k$, k vectores aleatorios independientes donde Y_j es la variable respuesta y X_j la covariable. Consideremos los modelos de regresión no paramétricos

$$Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j,$$

donde $\varepsilon_j \perp X_j$, $m_j(X_j) = E(Y_j|X_j)$ y $\sigma_j^2(X_j) = Var(Y_j|X_j)$.

En este caso se busca contrastar la hipótesis de igualdad de curvas de regresión,

$$H_0 : m_1 = m_2 = \dots = m_k \quad \text{frente a} \quad H_1 : m_i \neq m_j, \text{ para algún } i, j \in \{1, 2, \dots, k\}.$$

Pardo-Fernández, Van Keilegom y González-Manteiga (2007) proponen emplear estadísticos de contraste basados en las distancias entre las distribuciones empíricas de los residuos y las distribuciones empíricas de los residuos estimados bajo la hipótesis nula de igualdad de curvas de regresión. Para ello consideran las dos estimaciones de los errores ε_{ij} : $\hat{\varepsilon}_{ij0} = (Y_{ij} - \hat{m}(X_{ij}))/\hat{\sigma}(X_{ij})$ cuando la hipótesis nula es cierta y $\hat{\varepsilon}_{ij} = (Y_{ij} - \hat{m}_j(X_{ij}))/\hat{\sigma}(X_{ij})$ bajo la alternativa, siendo $\hat{m}_j(x)$ un estimador tipo núcleo de la función de regresión $m_j(x)$, $\hat{m}(x)$ un estimador de la función de regresión común $m(x) = m_1(x) = m_2(x) = \dots = m_k(x)$ y $\hat{\sigma}_j(x)$ un estimador de la función de varianza $\sigma_j(x)$. El estadístico de contraste se calcula a partir del proceso k -dimensional $\hat{\mathbf{W}}(y) = (\hat{W}_1(y), \dots, \hat{W}_k(y))^t$, $-\infty < y < \infty$, con $\hat{W}_j(y) = n_j^{1/2} (\hat{F}_{\varepsilon_{j0}}(y) - \hat{F}_{\varepsilon_j}(y))$, donde para cada $j = 1, \dots, k$, $\hat{F}_{\varepsilon_j}(y)$ es la distribución empírica de los residuos estimados bajo la hipótesis alternativa y $\hat{F}_{\varepsilon_{j0}}(y)$ la de los residuos estimados bajo la hipótesis nula. Así, el estadístico de tipo Kolmogorov-Smirnov basado en las distribuciones empíricas de los residuos es de la forma

$$T_{KS} = \sum_{j=1}^k \sup_y |\hat{W}_j(y)|.$$

De forma análoga al caso de bondad de ajuste de la sección anterior, si se supone que ε_j son variables continuas se puede obtener una estimación suavizada de F_{ε_j} y $F_{\varepsilon_{j0}}$ mediante un estimador tipo núcleo, es decir

$$\hat{F}_{\varepsilon_j}^{h_j}(y) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^y \frac{1}{h_j} k\left(\frac{s - \hat{\varepsilon}_{ij}}{h_j}\right) ds = \frac{1}{n} \sum_{i=1}^n K\left(\frac{y - \hat{\varepsilon}_{ij}}{h_j}\right)$$

y

$$\hat{F}_{\varepsilon_{j0}}^{h_{j0}}(y) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^y \frac{1}{h_{j0}} k\left(\frac{s - \hat{\varepsilon}_{ij0}}{h_{j0}}\right) ds = \frac{1}{n} \sum_{i=1}^n K\left(\frac{y - \hat{\varepsilon}_{ij0}}{h_{j0}}\right),$$

donde h_j y h_{j0} son las ventanas para suavizar las distribuciones empíricas de los residuos bajo la hipótesis alternativa y bajo la hipótesis nula, respectivamente.

A partir de las estimaciones suavizadas de las funciones de distribución se define el proceso $\hat{W}^s(y) = (\hat{W}_1^s(y), \dots, \hat{W}_k^s(y))$, con $\hat{W}_j^s(y) = n_j^{1/2} (\hat{F}_{\varepsilon_{j0}}^{h_{j0}}(y) - \hat{F}_{\varepsilon_j}^{h_j}(y))$. Luego, el estadístico de contraste de tipo Kolmogorov-Smirnov suavizado resulta

$$T_{KS,s} = \sum_{j=1}^k \sup_y |\hat{W}_j^s(y)|.$$

La selección de los parámetros ventana necesarios para el cálculo del estadístico de contraste se puede llevar a cabo minimizando una función de validación cruzada análoga a (2).

Se ha realizado un estudio de simulación para comparar el funcionamiento de ambos estadísticos a la hora de contrastar la igualdad de dos curvas de regresión. Al igual que en el caso del contraste para la bondad de ajuste las distribuciones de los estadísticos se han aproximado mediante un bootstrap suavizado de los residuos.

Para $i = 1, \dots, n$ se simulan muestras de dos poblaciones como $Y_{ij} = m_j(X_{ij}) + \sigma_j(X_{ij})\varepsilon_{ij}$, donde, para $j = 1, 2$, la distribución de ε_j es la normal estándar y X_j sigue una distribución uniforme en el intervalo unidad. El objetivo es contrastar $H_0 : m_1(x) = m_2(x)$, donde $m_1(x) = \exp(x)$ y $m_2(x) = \exp(x) + \delta x$, $\delta \in \{0, 0.1, 0.2, \dots\}$. Así, si $\delta = 0$ estaremos bajo la hipótesis nula, mientras que la desviación con respecto a H_0 aumenta a medida que δ crece. Se ha considerado una situación de homocedasticidad donde $\sigma_1(x) = \sqrt{0.25}$ y $\sigma_2(x) = \sqrt{0.5}$.

La Figura 2 muestra la proporción de rechazos en 1000 muestras de tamaños $(n_1, n_2) = (50, 50)$ y $(100, 100)$ para distintos valores de δ . El número de réplicas bootstrap es 200. De nuevo se observa una mejora en la potencia para todos los casos, aunque este fenómeno no es tan evidente como en el caso del contraste de bondad de ajuste.

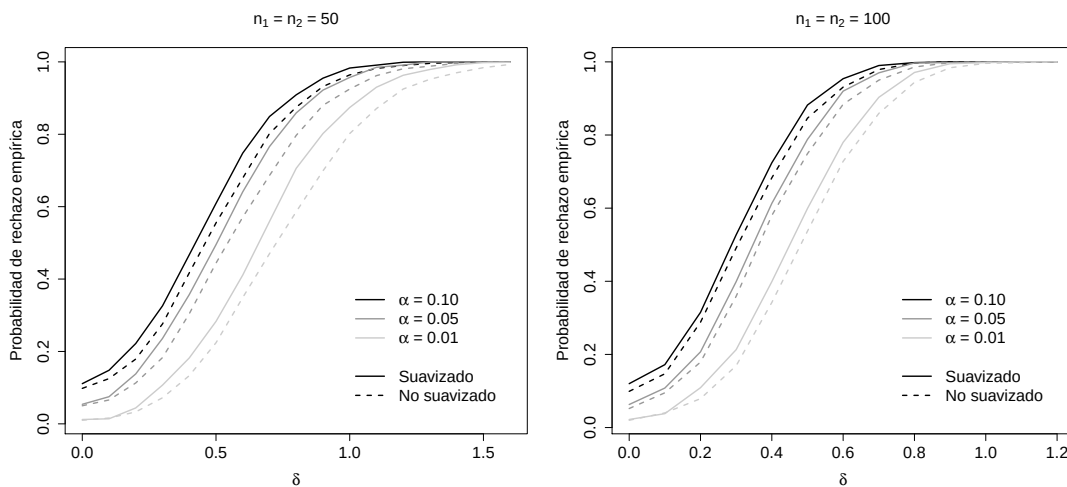


Figura 2: Curvas de potencia para el contraste de comparación de curvas de regresión.

REFERENCIAS

- Einmahl, J. H. J. y Van Keilegom, I. (2008) Specification tests in nonparametric regression. *Journal of Econometrics*, 143, 88–102.
- Pardo-Fernández, J. C., Van Keilegom, I. y González-Manteiga, W. (2007) Testing for the equality of k regression curves. *Statistica Sinica*, 17, 1115–1137.
- Racine, J. y Van Keilegom, I. (2019) A smooth nonparametric, multivariate, mixed-data location-scale test. *Journal of Business & Economic Statistics*. <https://doi.org/10.1080/07350015.2019.1574227>.
- Van Keilegom, I., González-Manteiga, W. y Sánchez-Sellero, C. (2008) Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST*, 17, 401–415.

XIV Congreso Galego de Estatística e Investigación de Operacións
Vigo, 24, 25 e 26 de outubro de 2019

HYPERSCALE 3D POINT CLASSIFICATION THROUGH FUNCTIONAL DATA ANALYSIS

Manuel Oviedo¹, Javier Roca², Carlos Cabo³, Celestino Ordóñez³

¹Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela (Spain)

²Department of Statistics and Operations Research, University of Vigo (Spain)

³Department of Mining Exploitation and Prospecting, University of Oviedo (Spain)

ABSTRACT

3D point clouds are collections of million or billion point coordinates that allows representing and modelling spatial objects. As they are not organized structures it is essential to develop algorithms to process the point clouds in order to extract useful information. One of the most popular and efficient approaches for 3D point cloud supervised classification is the use of multiscale feature extraction as predictors in the classifiers. Multiscale analysis outperforms single scale methods because it captures the local geometry that characterizes the objects at different scales. However, one of the main problems associated to multiscale feature extraction is how to determine the range and the number of scales, as well as to find out the most important scales. In this work, we analyze the use of functional data classification to solve these problems. The extraction of local features at very close scales allows to treat the data as functions and take advantage of the particularities of functional data analysis. The proposed methodology is applied to a real dataset and the results compared with those obtained using a conventional multiscale approach.

Keywords: 3D point cloud, laser scanning, photogrammetry, classification, multiscale, functional data analysis.

REFERENCES

- Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1-23.
- Demantke, J., Mallet, C., David, N. and Vallet (2011). Dimensionality based scale selection in 3d lidar point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38, Part 5: W12.
- Huang, R., Hong D., Hu, Y., Yao, W. and Stilla, U. (2019). Multi-Scale Local Context Embedding for LiDAR Point Cloud Classification. *IEEE Geoscience and Remote Sensing Letters*, 1-5, DOI: 10.1109/LGRS.2019.2927779.
- Weinmann, M. Jutzi, B. and Mallet, C. (2017). Geometric features and their relevance for 3d point cloud *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 157(4), 98?11.

